MATHEMATICAL Methods of Cybernetics

Yu. KORSHUNOV

Mir Publishers Moscow

MATHEMATICAL Methods of Cybernetics

Yu. KORSHUNOV

Mathematical Methods of Cybernetics

Ю. М. Коршунов

Математические основы кибернетики

Москва Энергоатомиздат

Yu. Korshunov

MATHEMATICAL METHODS OF Cybernetics

Mir Publishers Moscow

First published 1990
Revised from the 1987 Russian edition

Translated from the Russian by
M. EDELEV

УДК 681.5.01 : 51=20

Учебное издание

Юрий Михайлович Коршунов
МАТЕМАТИЧЕСКИЕ ОСНОВЫ КИБЕРНЕТИКИ

Заведующий редакцией В. К. Чугунов
Научный редактор Е. Г. Перова
Редактор Е. В. Кромина
Художник В. А. Медников
Художественный редактор А. Я. Мусин
Технический редактор Е. С. Потапенкова
Корректор С. В. Долженкова

ИБ № 8354

Сдано в набор 23.04.90. Подписано к печати 21.07.90. Формат 60×90/16.
Бумага типографская № 1. Печать высокая. Гарнитура обыкновенная.
Объем 12,50 бум. л. Усл. печ. л. 25,0. Усл. кр.-отт. 25,76.
Уч.-изд. л. 24,10. Изд. № 12/6694. Тираж 6200 экз. Зак. 01342.
Цена 2 р. 30 к.

Издательство «Мир» В/О «Совэкспорткнига»
Государственного комитета СССР по печати
129820, Москва, И-110, ГСП, 1-й Рижский пер., 2

Ордена Трудового Красного Знамени
Московская типография № 7 «Искра революции»
В/О «Совэкспорткнига» Государственного комитета СССР по печати
103001, Москва, Трехпрудный пер., 9

К $\frac{1602110000-326}{056(01)-90}$ 9—90

На английском языке

Printed in the Union of Soviet Socialist Republics

ISBN 5-03-001588-4

© «Энергия», 1980
© Энергоатомиздат, 1987 с изменениями
© English translation, M. Edelev, 1990

CONTENTS

Preface 7
List of Symbols 9

1. **INTRODUCTION** 11

1.1 Cybernetics Defined 11
1.2 Data Coding and Transmission 14
1.3 The Control System Defined 16
1.4 Complex Systems 20
Exercises 21

2. **BASIC CONCEPTS OF SET THEORY** 22

2.1 Finite and Infinite Sets 22
2.2 Set Algebra 29
2.3 Orderings and the Cartesian Product 37
2.4 Relations 40
2.5 Mappings and Functions 42
2.6 Relations Revisited 51
Exercises 54

3. **FUNDAMENTALS OF GRAPH THEORY** 55

3.1 Basic Concepts and Definitions 55
3.2 The Shortest Path Problem 63
3.3 Flows in a Transport Network 69

4. **ELEMENTS OF LINEAR ALGEBRA AND CONVEX SETS** 79

4.1 Metric Spaces and Distances 79
4.2 Metric Spaces in Cybernetics Problems 82
4.3 Linear Spaces 89
4.4 Euclidean Spaces 93
4.5 Matrices and Linear Transformations 97
4.6 Square Matrices 101
4.7 Symmetric Matrices and Quadratic Forms 110
4.8 Convex Sets 114
Exercises 122

5. **PROBABILITY THEORY AND MATHEMATICAL STATISTICS** 123

5.1 Probability Defined 123
5.2 Conditional Probabilities 128
5.3 Continuous Random Variables and Their Probability Distribution Functions 131
5.4 Expected Values and Moments 136

5.5 Two-Dimensional Random Variables 141
5.6 The Multivariate Normal Distribution 149
5.7 Stochastic Processes in Discrete Time 153
5.8 Basic Methods of Statistics 168
5.9 Regression Analysis and Experimental Design 185
Exercises 198

6. **OPTIMAL CONTROL PROBLEMS** 199

6.1 Basic Features of a Control Process 199
6.2 Optimization of Control Processes 202
6.3 Mathematical Description of the Controlled System 204
6.4 Classification of Optimal Control Problems 217
6.5 Sequential Control 223
6.6 Basic Concepts of Optimization 227

7. **LINEAR AND NONLINEAR PROGRAMMING** 235

7.1 LP Problem Formulation 235
7.2 Solution of LP Problems 243
7.3 Nonlinear Programming 252
7.4 Iterative Optimization Techniques 259
Exercises 266

8. **DYNAMIC PROGRAMMING** 267

8.1. Optimal Control as Variational Problem 267
8.2 Multi-Stage Decision Problems 271
Exercises 283

9. **GAME THEORY** 285

9.1 Introduction 285
9.2 Values of Games and Optimal Strategies 293
9.3 The Fundamental Theorem of Game Theory 299
9.4 General Techniques for Solving Games 306
Exercises 313

10. **STATISTICAL DECISION THEORY** 314

10.1 Statistical Games 314
10.2 Statistical Games Without Sample Information 318
10.3 Statistical Games with Sample Information 324
10.4 Use of Posterior Probabilities 330
10.5 Sequential Decision Analysis 336
10.6 Parameter Estimation and Filtering 344
Exercises 360

11. **SCHEDULING AND COMBINATORIAL PROBLEMS. THEORY OF QUEUES** 361

11.1 Scheduling and Sequencing 361
11.2 Project Planning and Control 363
11.3 Combinatorial Scheduling Problems 374
11.4 Examples of Problems Reducible to the Traveling Salesman Problem 377
11.5 Solution of Traveling Salesman Problems 379
11.6. Theory of Queues 386
References 396
Subject Index 399

PREFACE

Cybernetics is generally defined as a science dealing with regularities of control in various physical systems [1-6]. Many disciplines involving problems of control make wide use of the optimization techniques based on linear, nonlinear, and dynamic programming, game theory, statistical decision making, experimental design, and queueing theory. Many of these methods evolve from the general mathematical concepts of set theory, theory of graphs, vector spaces, probability theory and mathematical statistics. These common fundamentals of the many optimization techniques allow them to be covered compactly and from a common standpoint. This approach is appealing in that it simplifies the study of the aforementioned disciplines and renders their interconnections readily traceable.

The first chapter introduces the reader to the main concepts of control, cybernetics, and complex systems. In addition to the traditional mathematical tools of calculus, which are assumed known to the reader, the description of discrete complex systems requires special mathematical methods. These are sets, theory of graphs and linear algebra outlined in Chapters 2 through 4.

Chapter 5 discusses the methods of probability theory, statistics, regression analysis, and experimental design which are involved in simulating the stochastic behaviour of complex systems and uncertainty of the environment where they operate.

Whereas the early chapters deal with the construction of system models, Chapter 6 comes over to discussing the process of control. It gives a classification of solvable control problems, displays examples of mathematical model construction, and presents the concepts of optimization theory fundamental to solving the control problems in the following chapters.

Chapter 7 is devoted to the linear and nonlinear programming methods of solving constrained optimization problems. Such problems include many situations of managerial economics and process industries where an optimal solution in a multialternative environment is essential.

Chapter 8 devoted to dynamic programming takes up the methods of evaluating sequences of optimal control in processes which require

sequential decision making in search for an optimum or which are representable as sequential decision processes.

Chapters 9 and 10 consider the methods of decision making under conditions when the situation is not in full command of the decision maker. Some of the situations where "players" with opposite strategies are involved fall in the realm of game theory, other situations due to uncertainty of problem formulation are treated by statistical decision theory. The latter problems include signal estimation where no accurate measurements can be done, i.e. where signals are recovered from noisy environment.

Chapter 11 discusses a group of problems falling under scheduling problems and queueing theory. These problems arise in handling scheduling and sequencing tasks and industrial and social services control in real time.

All the problems discussed in this text are reduced to computer algorithms. However, to provide a better insight into the computational aspects of the methods discussed, the relevant illustrative examples are taken sufficiently simple to perform the necessary computations with merely pencil and paper.

The step-by-step presentation of the book allows for gradual increase of complexity of new concepts so that the material of preceding chapters serves as a basis for the following. It is assumed therefore that the reader will work through the book chapterwise. It does not mean, however, that individual chapters cannot be treated, if necessary, independently of the previous material.

Wherever conceptual or terminological difficulties may arise, the list of symbols and subject index can be of help in locating their origin in the text.

Although the book was written as an aid for students majoring in systems optimization and automatic control, it will be useful for students of other specialties with lesser weight on control but touching upon the concepts of optimization from the managerial standpoint as in scheduling and sequencing, production planning, project selection, cost-benefit analysis, and civil-sector problems.

Yu. M. Korshunov

LIST OF SYMBOLS

$\{\ldots\}$	set
$\in$	belongs to set
$\notin$	does not belong to set
$\subseteq$	inclusion symbol
$\subset$	belongs to set as subset
$\cup$	union
$\cap$	intersection
$\setminus$	difference of sets
$\times$	direct product of sets
$\bar{X}$	complement of set X
X^s	set X of order s
$\varnothing$	empty set
U	universe of discourse, universal set
R	set of all real numbers
M	system of sets
sup M	upper bound of set M
inf M	lower bound of set M
$\mathrm{Pr}_i M$	projection of set on axis i
$(a_1, \ldots, a_n)$	ordered set, vector
Λ	ordered empty set
$\forall$	for all
$\rightleftarrows$	equivalent by definition
$\rightarrow$	consequence, map
$f \circ g$	compound function of f and g and connective $\circ$
$\equiv$	equivalent
$\leqslant$	ordering symbol
$<$	symbol of strong order
$\gg$	symbol of dominance relation
$d(x, y)$	distance between two elements of set
$\| \mathbf{x} \|$	norm of x
E^n	n-dimensional Euclidean space
$C_{[a, b]}$	space of continuous functions

$\mathbf{x}$, $\mathbf{y}$	vectors
$\mathbf{A}$, $\mathbf{B}$	matrices
$\mathbf{x}^T$, $\mathbf{A}^T$	transposed vector and matrix
det $\mathbf{A}$	determinant of matrix $\mathbf{A}$
Z	sample space of outcomes of experiment
z	outcome of experiment, random variable
$\mathring{z}$	(central) random variable of zero mean
$p(z)$, $z \in Z$	probability distribution on space Z
$P(S)$	probability of event S
$p(z \mid S)$	conditional probability distribution
$p(y, z)$	joint probability distribution density
$\bar{z}$, $E(z)$, μ_z	average value, mathematical expectation, mean
σ, σ_z	mean square error, standard deviation
σ^2, σ_z^2	variance
$w(z)$	probability density function (p.d.f.)
$F(z)$	cumulative probability distribution function (c.d.f.)
$R(\mu, \omega)$	uniform distribution
$N(\mu, \sigma^2)$	normal (Gaussian) distribution
$w(r, n, p)$	binomial distribution
$w(r, a)$	Poisson distribution
$\binom{n}{m}$	permutations of n things m at a time
u	control
θ	state of nature
x	state of system
$T(x, u)$	transformation of system state
$q(x, u)$, $Q(x, u)$	objective function, performance measure
$J(u)$, $J_n(u)$	performance criterion, index
ξ, η	mixed strategies
$L(\xi, \eta)$	loss function
A	solution space
a	solution, action
a^*	Baesian solution
$R^*(\xi)$	losses incurred by Baesian decision
$d(z)$	decision function
$\rho(\theta, d)$	Baesian risk
$\rho^*(\xi)$	risk function
int x	integer part of x

1 INTRODUCTION

1.1 Cybernetics Defined

The ancient Greeks called cybernetics the art of navigation. The outstanding French mathematician and physicist André Ampére used this term to denote the science of governing a state. The modern meaning of this term is due to Norbert Wiener whose book *Cybernetics or Control and Communication in the Animal and the Machine* [1] pioneered this discipline.

The early machines need not much control. They were mainly engines, motors and simple actuators. Efficiency was the major performance index of such machines. Control was effected by the human operator in charge of the machine. He got the information necessary for the control by sensing or inspecting how the machine operates.

The progress of technology in the middle of this century led to the development of technical systems so sophisticated that a human operator was no longer able to control them properly. At the end of World War II one of such systems was to aim the airguns. At a speed of aircraft comparable with that of air shells, the system was to automatically track the path of the aircraft, calculate its probable route and correctly set the gun. Such systems had to rapidly collect information on the environment where the system operates and to extract from it the data necessary for system control. Saying it another way, such systems called for facilities necessary for communication and control. Fulfilment of this task rapidly advanced communication theory, computing, and automatic control.

The principal difference between the means of communication and control and the early machines was that their energy performance was no longer essential, instead the major attention was attracted to their ability to transmit and handle large amounts of data without distortion. For example, in a radio communication link only a small proportion of energy emitted by the transmitter reaches the receiver, consequently the efficiency of the radio link becomes excessively low. However, this link will be recognized as good if it conveys messages without large distortions and is immune to noise. Therefore, the main processes in the devices of communications

and control are those of transmission and processing of information, rather than those involved in energy conversion and utilization.

Because energy performance problems are inessential for communications and control, the analyst while handling a particular problem may detract from the physical features of information carriers and the physical nature of systems concerned and resort to general relations. Such general theory of communications and control applicable to any system independent of its physical origin is referred to as cybernetics.

The concept of a *system* and the notion of *control* are fundamental to cybernetics. Any actual system consists of many objects. These may be plants, human operators operating these plants, material resources and such. These objects are connected to one another and to the environment by certain links that represent forces, flows of energy, material, and information. Cybernetics detracts from the physical contents of the objects and their links and treats any actual system as an abstract set of elements endowed with some general properties and connected to one another by some relations defined by the nature of the links. This representation avoids the common categorization of systems into mechanical, electrical, chemical, biological and so on and reduces the system to a representation by an ensemble of interlinked and interacting elements. By way of example, Fig. 1.1 shows a cybernetic system constituted by four elements and six links.

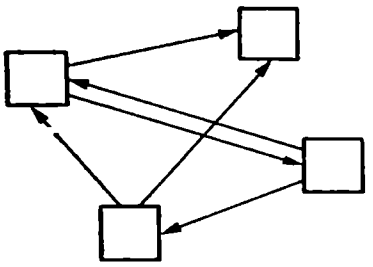


Fig. 1.1. Example of a cybernetic system

A system represented by a collection of elements may be conveniently described by the mathematical formalism of set theory that will be considered in the early chapters of this text. In many a system representation, the interelemental links can be described and analyzed by the methods of graph theory, and the interaction of elements is analyzed by the methods of linear algebra.

Practical systems may be broadly divided into deterministic and stochastic, according to the system structure and the nature of system links. A *deterministic system* is one in which the laws of motion are exactly known and the behavior can be forecast. For a *stochastic*, or *probabilistic*, system no forecasting of its behavior can be done. An example of deterministic system may be a clock. Such systems as production quality control, arrival of ships in ports, or movement of stocks in a warehouse with a number of suppliers and consumers are, obviously, of the probabilistic variety.

Problems handled by cybernetics lead as a rule to rather sophisticated probabilistic systems which involve a large number of ele-

ments with many diversified links. Most of industrial systems, and problems in economical, social, and biological areas produce such complex systems. Their mathematical description draws heavily on probability theory and mathematical statistics in addition to set theory and logics.

Up to this point we have touched on only the mathematical methods used for the description of cybernetic systems. Here we should recall that cybernetics deals with system control. Therefore we need to define the concept of control.

In a wide sense, the word *control* means guidance of some activity with the aim to achieve certain objectives. The process of control, or management, consists in making decisions as to the most appropriate actions in the situation at hand. The person in charge, referred to as the decision maker (DM), makes decisions upon analysis of the environment with the aid of the information he receives from his sense organs, measuring instrumentation, and other operators. In many a situation this information turns out insufficient for an unambiguous assessment of the system state. Then the DM resorts to his experience, memory, and intuition. An excellent property of the DM is his ability to make decisions under considerable uncertainty about the situation. Under the conditions of large industrial enterprises, however, even an experienced manager finds that his knowledge and intuition are insufficient for efficient control. There exist a number of mathematical methods, here collectively referred to as cybernetics and often combined under the name "operations research", that have been developed to unburden the DM of the activities associated with the collection and processing of information. These activities along with the analysis of feasible situations and preparation of recommendations for the DM are relegated to a computer. Such automatic systems are known as management information systems.

The role of computers deserves special mention. Modern high speed computers give the hand to the analyst in analyzing and comparing not only certain but all feasible solutions to the given problem. In so doing the computer may be programmed to select the most appropriate solution, i.e. solve the optimization problem.

In an industrial environment, a computer can easily handle a large amount of data concerning the production process and becomes a helpful aid of the manager.

For all the versatility of the computer, it must be equipped with certain mathematical techniques enabling it to analyze available information, screen it to extract the essential part to be used in the analysis of the current situation, and issue recommendations for the DM. The supporting mathematical disciplines include theory of information, game theory, desicion theory, queueing theory, linear programming and dynamic programming. Some of these new mathe-

matical methods will be discussed in this book. Selection of the methods for the text has been dictated by their significance and the space considerations of this volume.

1.2 Data Coding and Transmission

Communication links established between elements in a system may serve various purposes, for example, transmit energy, material, force, etc. In systems set up for the purpose of control of a plant, or management of an enterprise, we shall be mostly interested in the information capacity of the links, i.e. in their ability to convey information on the status of a system element.

Any information on some event inside the system or outside it is commonly called a *message*. Information links used to convey messag-

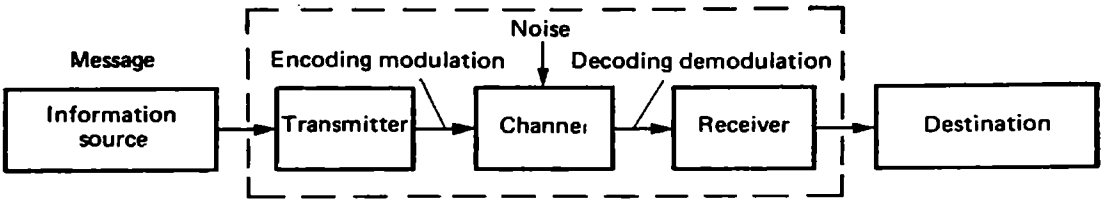


Fig. 1.2. Block diagram of a communication channel

es are referred to as *communication channels*. The physical carriers of information in the channels are called *signals*. The block diagram of a communication channel including data source and sink is shown in Fig. 1.2.

The word "signal" is from the Latin *signum*, i.e. a symbol. Accordingly, signals may be electrical, sound, visible, or other symbols employed to convey messages. A specific type of signals is a matter of agreement between the sides transmitting and receiving information and has nothing to do with the content of the transmitted information. Signals may be converted from one form into another without changing their information content, and can be stored, say in alphabetic form, for further use.

By way of example consider the type of signals employed in wiring a telegram. The initial message in alphanumeric form is converted by typing into a Morse code format, i.e. a sequence of dots and dashes. These are transmitted by wire as short and long impulses of electric current. At the receiving end the original text is recovered from the arriving pulse sequence. Conversion of signals from one form into another, as from alphanumeric form into a Morse code message as above, is called *encoding*. Any method of encoding involves a set of symbols, say letters, words, dots, etc., that has sense for both sending

and receiving sides. This set of symbols is known as *alphabet*. The symbols are set up by the agreement of both sides concerned.

One of the ideas underlying communications theory is that the symbols constituting an alphabet cannot be indefinitely different. Accordingly, the number of distinct symbols used in message transmission is finite. In the English language, for instance, all possible messages are compiled of 26 letters.

When transmitting a message the operator draws from the alphabet one symbol after another, converts them into the respective signals and sends them via a communication link. In the communication link, the signals are corrupted by noise so that a signal detected at the receiving end has to be identified with one of the symbols in the alphabet, i.e. a decision is to be made by discarding all the symbols but one most closely resembling the waveform arrived. This problem may be a formidable one if signals are severely distorted by noise. Methods developed to overcome these difficulties constitute communications theory.

A variety of alphabets are in use in actual communications systems. The more popular of them owing to a number of advantages is the *binary alphabet* consisting of two symbols denoted conventionally by 0 and 1. In this alphabet, any message will be a sequence of zeros and ones, e.g. 100110100. It is sometimes useful not to think of the 0 and 1, called binary digits or bits, as numbers but only as a pair of arbitrary symbols. By elementary combinatorial theory, the total number of distinct messages composed each of m binary digits is 2^m. In particular, the English alphabet consisting of 26 letters may be represented by a system having 5 binary digits, whereas the Russian alphabet consisting of 33 letters cannot. In a system of 6 bits offering $2^6 = 64$ distinct combinations, there is enough states for the Russian alphabet with all the punctuation marks. Consequently, the binary alphabet is a useful vehicle to represent and transmit any letter message. It can be employed also to transmit numeric data, though then some convenient code is required.

In the conventional decimal system, numbers are written with decimal digits from the range 0, 1, . . ., 9. Each digit in a group representing some number has a value depending on the position in the group. For example, the number 395 means

$$3 \times 10^2 + 9 \times 10^1 + 5 \times 10^0$$

the number 10 being the *radix*, or *base*, of this representation.

Any number N may be written in a representation with another radix R (integer). The representation will have the form ... $d_3d_2d_1d_0$ with $0 \leqslant d_i < R$ giving

$$N = \ldots d_3R^3 + d_2R^2 + d_1R^1 + d_0R^0 \qquad (1.1)$$

d_i being the numbers of N. For example in the *octal notation*, or code, used in some computers, the number 395 will be

$$6 \cdot 8^2 + 1 \cdot 8^1 + 3 \cdot 8^0$$

that is, will be written as "octal 613".

In a binary notation system, numbers are represented as strings of 0's and 1's. With radix 2 any number from 0 to 15 can be represented by the binary digits grouped in fours to make the hexadecimal code, for example,

$$3 = 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0, \quad \text{i.e.} \quad 0011$$
$$5 = 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0, \quad \text{i.e.} \quad 0101$$
$$9 = 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0, \quad \text{i.e.} \quad 1001$$

Conventionally, zeros at high positions may be dropped, i.e., 3 and 5 may be written as 11 and 101, respectively. Binary encoded, the number

$$395 = 2^8 + 2^7 + 2^3 + 2^1 + 2^0$$

will be a group of nine bits, 110001011.

Long strings of binary digits are inconvenient to read. Thus for human use it is often convenient to group the binary digits into groups of four or three bits at a time. In the *binary coded decimal notation*, for example, each digit position of a decimal number is represented by four bits. For example the number 395 will be written as

$$\underbrace{0011}_{3}\ \underbrace{1001}_{9}\ \underbrace{0101}_{5}$$

This notation, however, requires longer bit blocks than a binary representation, that is, would occupy more space in, say, computer memory. From the economical standpoint, the octal code (base 8) is preferable. The same number 395 in octal code will be written as

$$\underbrace{110}_{6}\ \underbrace{001}_{1}\ \underbrace{011}_{3}$$

1.3 The Control System Defined

The definition of management and control given earlier allows for a broad meaning. One finds elements of control in steering a car, navigating an aircraft, operating a lathe, managing an enterprise, and governing a state. Control may be exercised not only on various machines and mechanisms but also on processes. Some processes are supported by humans. Consequently, people may also come under control. The control process may be exercised by a human or

by an automatic device. For complex system control, such an automatic device is built around a computer.

Although situations where control is required are many and diverse, all of them are well described by the block diagram common to all systems. Figure 1.3 shows such a block diagram, in which blocks represent the system elements and arrows represent the connections between these elements. This structure characterizes any control process independent of whether it is performed by a human or an automatic device.

The key elements in a control system are the *controlled system* (sometimes referred to as the *process* or *plant*) and the controlling

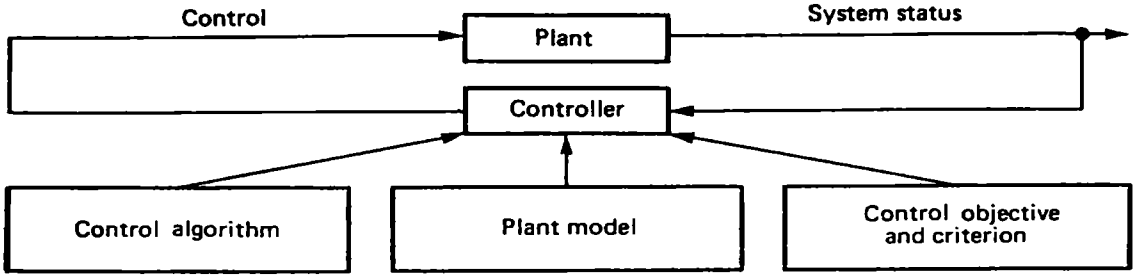


Fig. 1.3. Block diagram of a control process

system, or *controller*. In a simple example of a car driving process, the controlled system is the car and the controller functions are exercised by the driver. The controlling subsystem is connected with the process by a *feed-forward* path which normally comprises some forward controlling elements necessary to actuate the controlled system. In a car, these are steering gear, brakes, and transmission gear, which alter the character of motion. The perfection of these means is decisive for simplicity, flexibility, and reliability of control.

To judge the system behaviour, its output is measured or sensed and the result is fed back to the controller through the *feedback loop*. The amount and accuracy of information on the system state governs to a large extent the quality of control. Some systems are built as open-loop structures, i.e., without a feedback loop. However, this version is inferior to its feedback counterpart because of lower flexibility and efficiency.

The *objective* of control defines the state of the controlled system, its behaviour or motion that must be achieved as the result of control actions. Some *performance criteria* are devised to measure or judge how close to the set objective the controlling system has managed to drive the plant.

The term algorithm means a set of rules or procedures for the solution of a problem. A control algorithm prescribes how the technical

means of the feedforward path should be used depending on the control objective and process status. The necessary information about the process is supplied via the feedback loop.

The model of a controlled system is a collection of properties arranged into some mathematical form in the controlling system. The more detailed is this set of properties, the better is the control process. Frequently, however, the available knowledge of the process is insufficient for adequate control. Then the lacking properties have to be obtained immediately during the control process (adaptive control) or by trial and error.

The *mathematical model* of controlled system evolves as a rule in the form of a set of algebraic or differential equations. By solving

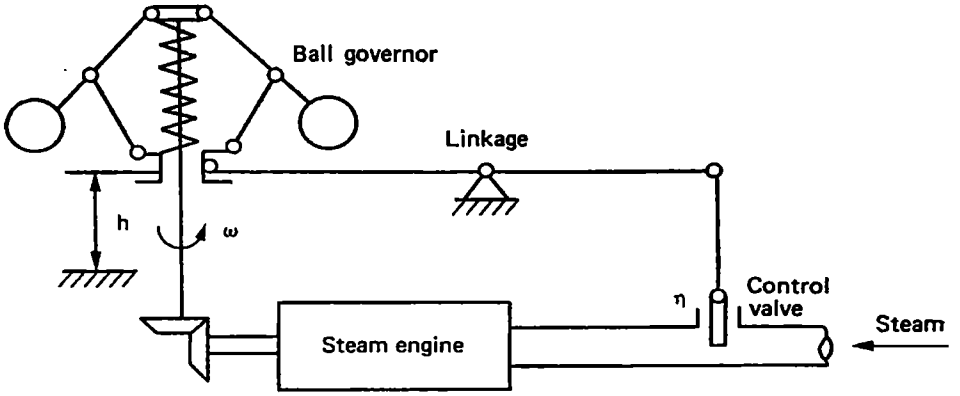


Fig. 1.4. Regulation of steam engine speed

these equations one establishes how the controlled system will behave at some or other stimulus signals impressed across its input to select a more appropriate method of control.

As an example of an automatic control system let us consider the speed stabilization system of a steam engine depicted in Fig. 1.4. The controlled system, or plant, is the steam engine. The shaft speed of the engine is varied by the control valve admitting steam to the engine. The controlling system is a centrifugal governor whose balls tend to move outwards and uplift the sleeve at higher shaft speeds.

Assume that the system is at equilibrium, i.e. the position of the governor balls and the valve is such that the shaft RPM ω is exactly equal to the setpoint ω_0. If under some external conditions such as variation of steam pressure, or variation of load, the equilibrium is upset, the system tends to restore the balance. For example, if the shaft speed has increased above the setpoint, the centrifugal governor balls uplift the sleeve and via the linkage shift the valve inside, thus decreasing the steam flow rate to the engine. A similar sequence of events, but with the reverse direction of movements, results if the shaft RPM drops below the setpoint ω_0.

A mathematical model of this system may be constructed as follows. The controlling system operates as soon as the shaft speed ω deviates from the setpoint ω_0, which implies a non-zero error

$$\varepsilon = \omega_0 - \omega \tag{1.2}$$

To simplify the discussion we assume that the system components have no inertia, i.e. the displacement of the governor sleeve, h, and the position of the valve, η, are proportional to the actuating error signal

$$\eta = k_1 \varepsilon \tag{1.3}$$

and the variation of ω is proportional to the position of the valve

$$\omega = k_2 \eta \tag{1.4}$$

The set of equations thus obtained is a mathematical model of the physical system. A graphical representation of such systems is

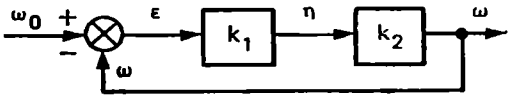


Fig. 1.5. Block diagram of the speed regulation system

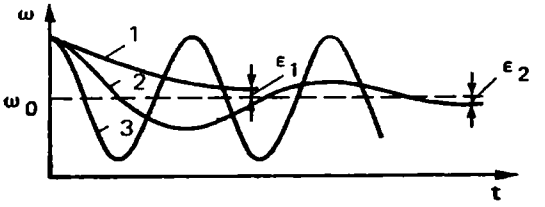


Fig. 1.6. Transient behavior of a system with gain increasing from (*1*) to (*3*)

by blocks showing the mathematical descriptions of the elements and lines connecting them. For the system under study, such a block diagram is given in Fig. 1.5.

A system model allows the designer to investigate the system behavior in advance and lay out ways for its optimum implementation. The measure of performance of the stabilization system is its accuracy governed by the value of error ε. The early systems of this kind developed as far back as two centuries could not offer a high accuracy and indeed were not subject to such a requirement. However, when progress in industry called for higher accuracy, designers were faced with formidable difficulties. The helping hand was given by mathematical modeling which related system error with the constructional parameters. Observing (1.3) in (1.4) and denoting $k = k_1 k_2$, called system gain, yields

$$\omega = k_1 k_2 \varepsilon = k\varepsilon$$

This expression for ω carries (1.2) to $\varepsilon = \omega_0 - k\varepsilon$, or

$$\varepsilon = \omega_0/(1 + k) \tag{1.5}$$

It can be seen that in this system the error cannot vanish and becomes sizeable at low k. This effect was observed by the designers of early speed governors.

Unfortunately attempts to raise the gain without bound to enhance accuracy lead to other undesirable effects associated with the nature of system transients. Figure 1.6 shows how the shaft speed approaches ω_0 with time at three distinct values of system gain k: low gain (1), high gain (2), and very high gain (3). At high gain, the processes in the system become oscillatory and the system gets unstable. This property also follows from the mathematical model if this allows for the elements inertia. The resolution of the incompatible requirements of high accuracy and stability is a key topic of control theory.

In analysis of systems as simple as that discussed above, there is no need for resorting to complicated methods of modeling. However, modern systems of automatic control are rather sophisticated and cannot be analyzed without mathematical methods outlined in this book.

1.4 Complex Systems

The rapid advance of science and technology in last decades has been accompanied by expansion of fields of research and evolution and sophistication of physical systems and processes. The modern systems of industrial enterprises, power supply, communication, transportation, and weaponry are so complicated that they are referred to as complex or large systems [8, 69].

Saying large systems we tend to emphasize not so much the physical dimensions of such systems as the variety of components constituting them, diversity of functions performed by the components, multiplicity of links connecting the components, and variability of external conditions affecting system's performance. At the same time, a large system is not a collection of elements, but rather an entity designed to perform a certain task. A large industrial enterprise, for example, includes a number of production facilities united into a single organism with warehouses, transportation, logistics, management and quality control.

Consistent operation of all parts in a large system cannot be achieved any longer with the methods of management used at the beginning of the century. The management and control of large systems draw heavily on automatic systems, above all those built around computers. Automatic control systems play the key role in managing large industrial plants. Computer assisted control system will be used in on-line process control, operation of flexible production lines, management, queueing system control, design and research activities.

EXERCISES

1.1. A system consists of seven elements each of which is connected to all others by links possessing two states, namely, connection allowed and connection blocked. What is the feasible number of system's states? Assume that we are to evaluate all states possessing property P with a time span of one second required to investigate one state; how much time does it take to evaluate all possible states?

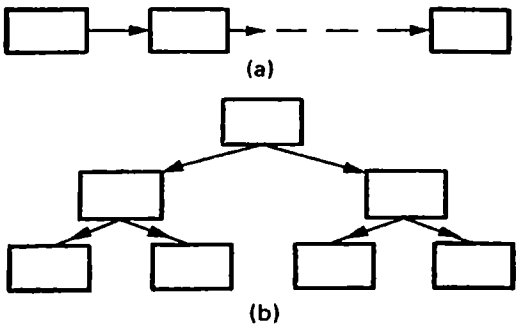


Fig. 1.7. Systems of (*a*) tandem and (*b*) hierarchical arrangement

1.2. Given the system of Problem 1.1 determine the reduction in the number of states if the system elements are connected by directed sequential links as in Fig. 1.7*a*, if the system has an hierarchical structure as in the diagram at (*b*). Give examples of systems having such types of connection.

1.3. The polynomial (1.1) used to represent a number in a notation with radix R may be written in the form

$$N = \{[(\ldots d_3)\, R + d_2]\, R + d_1\}\, R + d_0 \qquad (1.6)$$

which is equivalent to the sequence of expressions

$$N = N_1 R + d_0, \quad N_1 = N_2 R + d_1, \quad N_2 = N_3 R + d_2, \ \ldots$$

Use these formulae to verify the following rule of representing the decimal number 395 in octal notation

395 : 8 = 49 remainder 3 or in shorthand	395	
49 : 8 = 6 remainder 1	49	3
6 : 8 = 0 remainder 6	6	1
	0	6

This operation produces the octal number $(613)_8$.

Formulate a rule of representing decimal numbers in binary notation.

1.4. Give the binary, octal, and hexadecimal representations of the following decimal numbers: 27, 467, 519, and 1263.

Note: Use A, B, C, D, E, and F for the hexadecimal numbers 10 to 15.

1.5. Represent the numbers of Problem 1.4 in the binary coded decimal notation and in octal. How many binary digits do you need to write one digit in octal notation? Which of these representations is more economical in terms of bit requirement?

1.6. Consider whether or not the systems of traffic control with the aid of (1) a traffic light preset for a red-yellow-green cycle, 2) a traffic control officer have a feedback loop.

2

BASIC CONCEPTS OF SET THEORY

2.1 Finite and Infinite Sets

2.1.1 Fundamental Definitions

The concept of a *set* is a fundamental notion that avoids definition. Intuitively a set may be understood as a collection of distinct objects treated as a whole.

One may speak of a set of chairs in a room, people in a city, students in a group, positive integers, letters in an alphabet, states of a system, etc. However, the concept of a set is applicable only where the elements are distinct. For instance, one cannot speak of a set of drops in a glass of water since it is impossible to localize and indicate each particular drop in such a "set".

The objects constituting a set are referred to as *members*, *elements*, or *points* of the set. For example, the number 3 is an element of the set of natural numbers, and the letter b is an element of the set of the English alphabet.

Braces are commonly used to specify a set. We shall use capital letters A, S, X, . . . for sets and lowercase a, s, x, . . . for elements. The notation $a \in S$ means "a is an element of S", whereas $x \notin S$ means "x is not an element of S". Writing $x_1, x_2, \ldots, x_n \in S$ is a shorthand for $x_1 \in S, x_2 \in S, \ldots, x_n \in S$.

A set is said to be finite or infinite according as it contains a finite or infinite number of elements. Thus, for a finite set there exists a natural number N giving the number of elements in this set.

Two ways exist to specify a set, listing and description. For example, the set of students in a group may be specified by listing their surnames. To make the long record $X = \{x_1, x_2, \ldots, x_n\}$ short, we shall write $X = \{x_i\}_1^n$ or introduce the set of subscripts $I = \{1, 2, \ldots, n\}$ and write $X = \{x_i\}$, $i \in I$. This notation is convenient in dealing with finite sets which contain small number of elements, though it may occasionally be used in specifying infinite sets as $\{2, 4, 6, \ldots\}$. Naturally this notation is applicable in unambiguous circumstances when it is quite clear what the dots mean.

The descriptive method consists in specifying the characteristic property which all members of the set possess. If Z is the set of integers, then the set A of all natural numbers 0, 1, 2, . . . may be

specified as

$$A = \{x \in Z\colon x \text{ is a natural number}\}$$

and read as follows: set A consists of elements x of set Z for which the property of being a natural number is true.

In unambiguous situations where it is quite clear to which set elements x belong, no specification of x belonging to Z is necessary. The respective notation will be

$$A = \{x\colon x \text{ is a natural number}\}$$

Examples of specifying sets by means of description follow.

$\{x\colon x$ is an even number$\}$ is a set of even numbers

$\{x\colon x^2 - 1 = 0\}$ is a set $\{+1, -1\}$

Let C be a set of integers, then $\{x \in C\colon 0 < x \leqslant 7\}$ is the set $\{1, 2, 3, 4, 5, 6, 7\}$.

The *empty set*, or *null set*, denoted $\varnothing$, is the set with no elements. For example,

$$\{x \in C\colon x^2 - x + 1 = 0\} = \varnothing$$

This concept plays an important role in specifying sets by description. For example, we could not say of the set of real valued roots of the quadratic equation without previously verifying whether this equation has real valued roots. In other words this concept enables one to operate with the set of real valued roots with no worry about actual presence of real valued roots in the set under investigation. We shall conditionally think of empty sets as finite sets.

Let us now look at the notion of equal sets. Two sets are said to be *equal* if both consist of the same elements, i.e. represent one and the same set. Two sets X and Y are not equal, written $X \neq Y$, if either X has elements that do not belong to Y, or Y has elements that do not lie in X. The relation of equality is reflexive, symmetric, and transitive, namely:

reflexivity, for all $x \in X$ $x = x$;
symmetry, whenever $X = Y$, then also $Y = X$;
transitivity, if $X = Y$ and $Y = Z$, then $X = Z$.

From the definition of equality of sets it follows that the order of elements in the sets is immaterial. For example, the notations $\{3, 4, 5, 6\}$ and $\{4, 5, 6, 3\}$ represent one and the same set.

The concept of a set suggests that a set must not contain indistinguishable members. The notation $\{2, 2, 3, 5\}$ should be recognized as incorrect and replaced by $\{2, 3, 5\}$. This may imply the set of simple divisors for the number 60.

2.1.2 Subsets

A set X is a subset of a set Y, or is contained in Y, if every element of X is an element of Y. For instance $\{3, 4, 5\}$ is a subset of $\{1, 2, 3, 4, 5\}$ since each of the elements 3, 4, 5 of the first set belongs also to the second.

It is often convenient to write set-theoretical expressions in symbolic form. We shall use two such symbols, $\forall$ meaning "for all" and $\rightarrow$ meaning "implies", in defining subsets. For example, the phrase "for any x the assertion 'x belongs to X' implies 'x belongs to Y'" will be written as

$$\forall x \, [x \in X \rightarrow x \in Y] \tag{2.1}$$

When a set X is a subset of a set Y we write

$$X \subseteq Y \tag{2.2}$$

which reads as Y contains X. The symbol $\subseteq$ implies inclusion. Topologists often use the symbol of strong inclusion $\subset$ to emphasize that Y contains other elements in addition to the elements of X. Then

$$X \subset Y \tag{2.3}$$

Symbolically the connection between these two symbols may be given as

$$X \subset Y \rightleftarrows X \subseteq Y \quad \text{and} \quad X \neq Y \tag{2.4}$$

Here we used the symbol $\rightleftarrows$ implying equivalence.

It is instructive to note some properties of subsets following from the definition of a subset, namely,

reflexivity, $X \subseteq X$

transitivity, if $X \subseteq Y$ and $Y \subseteq Z$, then $X \subseteq Z$

By convention, for all sets M

$$\varnothing \subseteq M \tag{2.5}$$

The empty set $\varnothing$ does not contain elements, therefore, being added to M it does not change its elemental composition. For logical purposes we assume that any set M contains the empty set as a subset.

2.1.3 Correspondence Between Sets

Consider two sets, a flock of four sheep and a group of four trees. A relation that can be established between these sets cannot be established between either of them and a pile of five stones or a group of seven trees. Indeed, the former two sets may be put in correspondence, say by binding sheep to trees so that each sheep is

paired with exactly one tree. This relation is known as *one-to-one correspondence.*

Let X and Y be m- and n-element sets. They can be put into one-to-one correspondence if only $m = n$. How many such one-to-one correspondence relations exist for two n-element sets X and Y? Assume that these sets are ordered. Then the first element of X may be associated with any one of n elements in Y. For every such correspondence, the second element of X may be associated with any one of $n - 1$ remaining elements in Y, and so forth. Having established such a correspondence for $n - 1$ elements of X, the last element of this set may be put in correspondence with the sole element left in Y. Thus the total number of one-to-one correspondence relations for n-element sets will be

$$n\,(n-1)\ldots 1 = n!$$

Example 2.1. Determine the number of k-element subsets contained in an n-element set M.

By putting the set M into one-to-one correspondence with the set of natural numbers $N = 1, 2, \ldots, n$ we find that the number of versions in which we can establish such a correspondence is $n!$. Taking the first k elements of each version we obtain $n!$ subsets of interest. These, however, will include identical combinations of which only one should be considered in the calculation.

Let the first k elements of some correspondence form a set M_k'. The same set is formed by the first k elements in all correspondences that differ from the former by either the order of these elements, i.e. the total of $k!$, or the order of the last $n - k$ elements, i.e. the total of $(n - k)!$ Consequently, the total number of correspondences that form the k-element subset M_k' is $k!\,(n - k)!$ Any other k-element subset of M will form the same number of correspondences of the total number $n!$ Therefore the total number of k-element subsets contained in M will be

$$\frac{n!}{k!\,(n-k)!} = {}_nC_k = \binom{n}{k}$$

i.e. the number of combinations of n elements taken k at a time.

Now it is an easy matter to determine the total number L of all possible subsets of an n-element set M, namely,

$$L = \sum_{k=0}^{n} {}_nC_k = 2^n$$

Here, the numbers ${}_nC_0 = 1$ and ${}_nC_n = 1$ define the empty set $\varnothing$ and the set M itself. The sets $\varnothing$ and M are called *improper* subsets of M. All the other subsets are *proper* subsets of M.

2.1.4 Countable Sets

Trying to put infinite sets into one-to-one correspondence one would be faced with difficulties associated with the need to operate infinite numbers of elements. A way out is found by putting such sets in correspondence with the series of positive integers, 1, 2, . . ., n, ... A set of objects whose members can be put into one-to-one correspondence with the positive integers is called *countable*, or countably infinite. Such a correspondence exists not for all infinite sets, of course. The sets of all integers and of all rational numbers are countable, but the set of all real numbers is not.

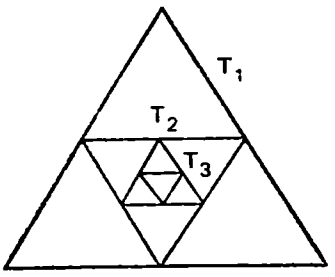


Fig. 2.1. Infinite set of equilateral triangles

By way of example consider an infinite set of equilateral triangles with the geometry evident from Fig. 2.1. This set can be put into one-to-one correspondence with the positive integers by ordering the triangles T according to decreasing size, i.e., T_1, T_2, etc. Consequently, this set of triangles is infinitely countable. There exist infinitely many other equilateral triangles that do not belong to the set under consideration. The problem of whether the set of all equilateral triangles, or all the triangles, is countable requires a separate investigation.

Some examples of countable sets follow.

1. The set of squares of integers, 1, 4, 9, . . ., n^2, . . ., is a subset of the set of natural numbers. However, the set is infinite countable since it can be put into one-to-one correspondence with the positive integers by assigning to each element a positive integer which being squared gives this element.

2. The set Z of all integers, positive and negative, is countable, although the natural numbers are the subset of this set. This fact can be established by considering the one-to-one correspondence

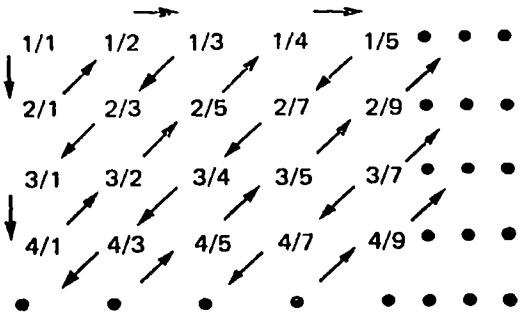
$$\begin{aligned} N &= 1\ 2\ \ 3\ \ 4\ \ 5\ 6\ \ 7\ \ldots \\ Z &= 0,1\ -1\ \ 2\ -2\ 3\ -3\ \ldots \end{aligned}$$

from which it follows that

$$\begin{aligned} z_n &= n/2, && \text{for even } n \\ &= -(n-1)/2, && \text{for odd } n \end{aligned}$$

3. The fact that the set of all rational numbers is countable is still more amazing. (Rational numbers are those representable as q/p, where p and q are any integers and $q \neq 0$.) To prove this fact let us represent the set of rational numbers in the form of a table

whose elements cannot be reduced any further



Moving through the table in the direction indicated by arrows we arrive at the sequence

$$1,\ 2,\ \frac{1}{2},\ \frac{1}{3},\ \frac{2}{3},\ 3,\ 4,\ \frac{3}{2},\ \frac{2}{5},\ \frac{1}{4},\ \frac{1}{5},\ \frac{2}{7},\ \frac{3}{4},\ \frac{4}{3}\ \ldots$$

whose elements may be numbered.

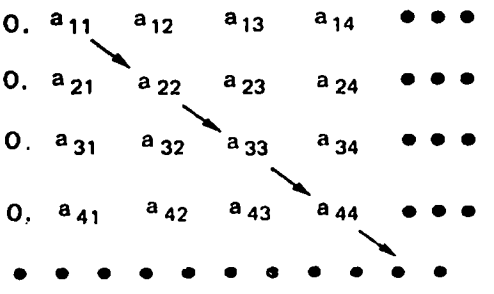
From the examples given above we may infer one of the outstanding properties of infinite sets, namely, that an infinite set may be put into one-to-one correspondence with its proper infinite subset. No such property exists for finite sets.

Existence of sets that are not countable follows from a theorem proved by the German set theorist G. Cantor in 1874.

Theorem 2.1 The set of all real numbers in the interval $0 < x \leqslant 1$ is not countable.

Notice that any number within the interval in question is a finite or infinite decimal fraction $0.\ a_1a_2a_3a_4\ \ldots$ and may be represented by a point of the interval of the real axis. The assertion of the theorem is, consequently, that the set of points of the interval $(0, 1]$ is uncountable.

Proof. Assume that the sequence $x_1, x_2, x_3, \ldots$ is an infinite list of real numbers that belong to this interval. We seek the answer to whether or not the list may contain all the numbers of this interval, or putting it another way, whether there exists a number which belongs to this interval but is not in the list. To find such a number we put down all the decimal fractions from the list one under another



Now we form a fraction whose digits are indicated by the arrow sequence and replace each of the digits a_{nn} with another digit a'_{nn} so that the resultant fraction should be infinite. This fraction $0.a'_{11}a'_{22}a'_{33}a'_{44}\ldots$ is a real number lying within the interval but not in the list. This fraction differs from the first of the fractions by its first digit following the decimal point, from the second fraction by its second digit, from the third by its third digit and so on.

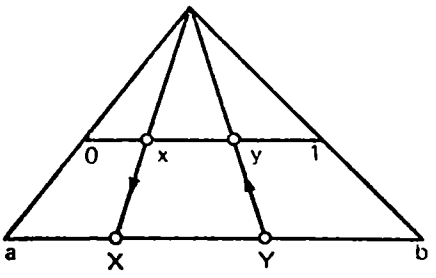


Fig. 2.2. One-to-one correspondence between interval (0, 1] and interval $(a, b]$

Unusual properties of uncountable sets manifest themselves in that the interval under investigation (0, 1] may be put into one-to-one correspondence with any other interval $(a, b]$, say by the central projection method as shown in Fig. 2.2. Thus the set of all real numbers in any interval $(a, b]$ is uncountable.

2.1.5 Supremum and Infimum of a Set

Considering a set of real numbers, one may compare its members by magnitude and, provided the set is finite, evaluate the smallest and largest elements of the set. For finite sets specified by listing, this is a rather easy task. For instance, in the set $T = \{4, 3, 5, 6\}$, $\max T = 6$ and $\min T = 3$. However, if the set is specified by description, say a rule is given on how to compute its elements, then the evaluation of the extremal element may be rather hard. A somewhat easier problem seems to be the evaluation of the region confining all the elements of a set. The concepts of upper and lower bounds of a set are quite useful in handling such problems.

Let S be a set of real numbers. An upper bound of S is a number C such that for any $x \in S$ we have $x \leqslant C$. Candidates for the upper bound of a set may be infinitely many, and may be not existent at all. For example, in the set $m < S < M$, any $C \geqslant M$ is an upper bound. The set of all integers has no upper bound.

If S is a set of real numbers bounded from above, then there is a smallest real number C such that $x \leqslant C$ for all x in S. The number C is called the least upper bound or *supremum* of S and denoted $\sup S$ or

$$\sup_{x \in S}(x) \quad \text{or} \quad \sup x\colon\ x \in S$$

In the example above, $\sup S = M$. A set may have only one least upper bound since if C_1 and C_2 are two such bounds, then $C_1 \leqslant C_2$ and $C_2 \leqslant C_1$ and, consequently, $C_1 = C_2$.

Similarly, the greatest lower bound or *infimum* of a set S is denoted inf S or

$$\inf(x) \quad \text{or} \quad \inf_{x \in S} x: \; x \in S$$

In the example above, inf $S = m$.

Theorem 2.2 If $B \subseteq A$, then

$$\inf B \geqslant \inf A \quad \text{and} \quad \sup B \leqslant \sup A \tag{2.6}$$

Proof. Denote by b' an element of B which is least in value, i.e., $b' \in B$ and $b' = \inf B$. Since $B \subseteq A$, $b' \in A$. Let a' be an element of A having the least value, i.e. $a' \in A$ and $a' = \inf A$. If $b' = a'$, then $b' = \inf A$, if $b' \neq a'$, then $b' > a' = \inf A$. Thus, $b' \geqslant \inf A$ or $\inf B \geqslant \inf A$.

The remaining part of the theorem may be proved by a similar argument.

2.2 Set Algebra

2.2.1 Preliminary Remarks

Here we have an intention to develop a kind of algebra of sets, that is, methods for combining sets to form other sets, together with some general laws that hold for these combinations. These laws are very similar to those of elementary algebra. We recall them below.

Let a and b be some numbers, $a + b$ be the sum, and ab be the product of these numbers. The laws of elementary algebra that hold for the sum and product are as follows.

1. Commutative law: $a + b = b + a$, $ab = ba$.
2. Associative law: $(a + b) + c = a + (b + c)$, $(ab)\, c = a\,(bc)$.
3. Distributive law: $(a + b)\, c = ac + bc$.

It is worth noting that both in associative and in commutative laws we may replace addition with multiplication and vice versa. The resultant law will hold as well as the original. No such symmetry exists in the distributive law. If we replaced addition by multiplication in this law and multiplication by addition, the result would be absurd

$$(ab) + c = (a + c)\,(b + c)$$

One may ask whether there exists an algebra where the distributive law would be as symmetric with respect to addition and multiplication as the commutative and associative laws are. Such an algebra happens to exist — this is the algebra of sets where all three laws are symmetric with respect to summation and multiplication.

Similarity between the operations of addition and multiplication manifests itself also in the existence of two remarkable numbers,

0 and 1, such that addition of the first and multiplication by the second do not alter any number, namely,

$$a + 0 = a, \quad a \cdot 1 = a$$

Notice that the last expression can be obtained from the former by substituting · for + and 1 for 0. However, the similarity between addition and multiplication does not extend much further. For instance, the number 0 plays a special role as compared with all other numbers including unity. This distinction of 0 follows from the relation $a \cdot 0 = 0$. If we replace multiplication by addition in this sentence, we arrive at $a + 1 = 1$, which is almost never true. We shall see that the similarity between zero and unity will be much more significant in set algebra than in classical algebra.

2.2.2 Unions

The union of X and Y, written $X \cup Y$, consists of all the elements of X together with those of Y, thus

$$X \cup Y = \{x \mid x \in X \quad \text{or} \quad x \in Y\} \tag{2.7}$$

The union of sets is sometimes referred to as the sum of sets and denoted $X + Y$. However the properties of the union somewhat differ from the properties of the sum in ordinary arithmetic context, therefore we shall not use this term in our further discussion.

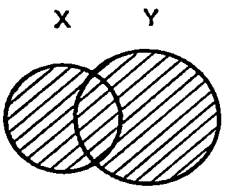


Fig. 2.3. Union of two sets

Example 2.2. If $X = \{1, 2, 3, 4, 5\}$ and $Y = \{2, 4, 6, 7\}$, then $X \cup Y = \{1, 2, 3, 4, 5, 6, 7\}$.

Example 2.3. Venn Diagram. Unions are often pictured using a Venn diagram which represents the sets concerned as regions in the plane, selected so that any general formula about sets which holds for the regions is automatically valid in general. For two sets X and Y represented by circles, the union $X \cup Y$ is given by the Venn diagram in Fig. 2.3 as hatched area confined by two circles.

The concept of the union may be extended to a larger number of sets. Denote by $M = \{X_1, \ldots, X_n\}$ an indexed family of sets. The union of these sets

$$\bigcup_{i=1}^{n} X_i = \bigcup_{X \in M} X = X_1 \cup \ldots \cup X_n \tag{2.8}$$

represents a set consisting of all those elements which belong to at least one of the sets of family M.

Using Venn diagrams one can easily prove that the commutative and associative laws are valid for unions of sets, namely,

$$X \cup Y = Y \cup X \tag{2.9}$$

$$(X \cup Y) \cup Z = X \cup (Y \cup Z) = X \cup Y \cup Z \tag{2.10}$$

Indeed, the right and left parts of these expressions consist of one and the same elements. Also,

$$X \cup \varnothing = X \tag{2.11}$$

This expression is self-evident since the empty set does not contain elements, hence X and $X \cup \varnothing$ consist of the same elements. Equation (2.11) suggests that the empty set plays the role of zero in set algebra.

2.2.3 Intersections

The intersection of two sets X and Y is the set whose elements are those objects belonging to both X and Y. It is denoted $X \cap Y$, and defined formally by

$$X \cap Y = \{x \mid x \in X \quad \text{and} \quad x \in Y\} \tag{2.12}$$

Example 2.4. For sets X and Y in Example 2.2 $X \cap Y = \{2, 4\}$.

Example 2.5. Venn Diagram. If X is the set of points of the left circle in Fig. 2.4*a* and Y is the set of points of the right circle, then $X \cap Y$ is represented by the hatched area common to both circles.

The sets X and Y are non-intersection or disjoint if they have no elements in common, i.e.

$$X \cap Y = \varnothing \tag{2.13}$$

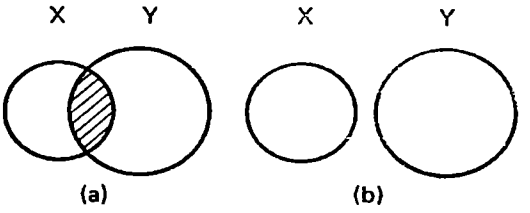


Fig. 2.4. (*a*) Intersection of two sets. (*b*) Nonoverlapping sets

Example 2.6. Examples of disjoint sets are

(1) the sets $\{1, 2, 3\}$ and $\{4, 5, 6\}$,

(2) sets of points of circles X and Y in Fig. 2.4*b*.

One more difference of set algebra from elementary algebra is worth noting. If a and b are two numbers, they can be related in either of three ways

$$a < b, \quad a = b, \quad b < a \tag{2.14}$$

For sets X and Y, however, neither of the relations may hold

$$X \subset Y, \quad X = Y, \quad Y \subset X \tag{2.15}$$

For example, if X is a set of postgraduate students, Y is a set of students who live in the campus, then the three above relations would

imply; $X \subset Y$, that every postgraduate student lives in the campus; $X = Y$, that the campus shelters postgraduate students only; and $Y \subset X$, that all students who live in the campus are postgraduates.

Obviously all these relations do not exhaust all possibilities. As follows from the previous definitions, X and Y may be combined by one of the five relations as follows:

$$X = Y, \quad X \subset Y, \quad Y \subset X, \quad X \cap Y = \varnothing$$

and X and Y have a nonzero intersection.

The concept of intersection may be extended to more than two sets. For an indexed family of sets $M = \{X_1, \ldots, X_n\}$, the intersection is written as

$$\bigcap_{X \in M} X = \bigcap_{i=1}^{n} X_i = X_1 \cap \ldots \cap X_n \tag{2.16}$$

and represents a set whose elements belong to each of the sets of the family M.

It is not hard to see that the intersection is commutative,

$$X \cap Y = Y \cap X \tag{2.17}$$

and associative

$$(X \cap Y) \cap Z = X \cap (Y \cap Z) = X \cap Y \cap Z \tag{2.18}$$

Also the relation

$$X \cap \varnothing = \varnothing \tag{2.19}$$

similar to the algebraic $a \cdot 0 = 0$ takes place. This relation along with (2.11) indicates that the empty set plays the role of zero in set algebra.

2.2.4 Difference

Unlike union and intersection, this operation is defined for two sets only. The difference $X \setminus Y$ of two sets X and Y is the set consisting of all those and only those elements which belong to X and do not belong to Y. The Venn diagram for the difference is shown in Fig. 2.5. Formally we may define the difference as

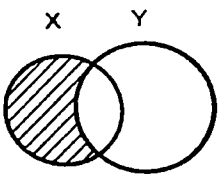


Fig. 2.5. Difference of two sets

$$X \setminus Y = \{x \mid x \in X \quad \text{and} \quad x \notin Y\} \tag{2.20}$$

Example 2.7. For sets X and Y of Example 2.2, $X \setminus Y = \{1, 3, 5\}$ and $Y \setminus X = \{6, 7\}$. For sets X and Y of Example 2.3 $X \setminus Y$ is presented by the hatched area in Fig. 2.5.

2.2.5 Universe of Discourse

We have learned in the earlier sections that the role of zero in set algebra is played by the empty set. In many mathematical discussions it would be helpful to have a set U which would play the role of unity, so that the relation

$$X \cap U = X \tag{2.21}$$

holds that is similar to the relation $a \cdot 1 = a$ of elementary algebra.

The relation (2.21) would imply that the intersection or the common part of the sets U and X coincide for any set X with this very set. This, however, is possible only when the set U contains all elements of which the set X can consist, so that any set X is completely contained in U. Such a set is termed a *universe of discourse.*

If in some particular discussion we include only subsets of some fixed set U, then this largest set U is the universe of discourse. The role of universe of discourse may be taken up by various sets. For instance in discussing sets of students in a group, good students, students with scholarship, students living in the campus, the role of a universe of discourse is played by the set of students in the group.

We shall represent a universe of discourse by a set of points in the form of rectangle in Venn diagrams. Areas of this rectangle represent subsets of the universe of discourse.

The universe of discourse possesses a remarkable property which has no analogies in common algebra, namely, for any set X,

$$X \cup U = U \tag{2.22}$$

Indeed, the union $X \cup U$ is a set including all elements of X and all elements of U. Observing that U includes all elements of X, we recognize that $X \cup U$ will consist of the same elements as U itself.

2.2.6 Complements

The set $\overline{X}$ defined as

$$\overline{X} = U \setminus X \tag{2.23}$$

is called the complement of X (relative to U) often suppressing the parenthesized phrase. In the Venn diagram of Fig. 2.6, the complement of $\overline{X}$ is represented by the open area. The formal definition of the complement of X is

$$\overline{X} = \{x \mid x \in U \text{ and } x \notin X\}$$

Example 2.8. If $U = \{1, 2, 3, 4, 5, 6, 7\}$ and $X = \{3, 5, 7\}$, then $\overline{X} = \{1, 2, 4, 6\}$.

From (2.23) it follows that X and $\bar{X}$ have no elements in common, so that

$$X \cap \bar{X} = \varnothing \tag{2.24}$$

There are no elements of U that do not lie in either X or $\bar{X}$, since those elements which do not belong to X lie in $\bar{X}$. Hence,

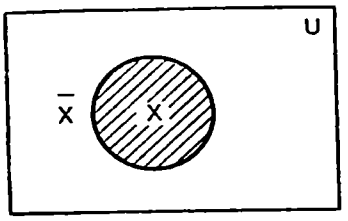


Fig. 2.6. Complement of a set

$$X \cup \bar{X} = U \tag{2.25}$$

From the symmetry of (2.24) and (2.25) with respect to X and $\bar{X}$ it follows that not only $\bar{X}$ is the complement of X, but also X is the complement of $\bar{X}$, i.e.

$$\bar{\bar{X}} = X \tag{2.26}$$

The complement is a convenient vehicle in representing the difference

$$X \setminus Y = \{x \mid x \in X \text{ and } x \notin Y\} = \{x: x \in X \text{ and } x \in \bar{Y}\}$$

that is,

$$X \setminus Y = X \cap \bar{Y} \tag{2.27}$$

2.2.7 Partitions

Partition is a frequent operation on sets resulting in a system of subsets. If N is the set of natural numbers, and A_0 and A_1 are the sets of even and odd numbers, then the family $\{A_0, A_1\}$ is a partition of the set N. This set may be partitioned also into a set of numbers divisible by 3 with zero remainder, with a remainder of 1 and with a remainder of 2.

In order to put the definition of partitioning on a firm footing, consider a set M and an indexed family $\mathrm{M} = \{X_1, \ldots, X_n\}$. The family is said to be a partition of M if

(1) any set X of M is a subset of M,

$$\forall X \in \mathrm{M}\ [X \subseteq M] \tag{2.28}$$

(2) any two sets X and Y of M are disjoint,

$$\forall X, Y \in \mathrm{M}\ [X \neq Y \rightarrow X \cap Y = \varnothing] \tag{2.29}$$

(3) the union of all sets constituting the partition gives M,

$$\bigcup_{X \in \mathrm{M}} X = M \tag{2.30}$$

We revisit the concept of partitioning in considering the relation of equivalence, a closely allied concept.

2.2.8 Some Useful Identities

The concepts of union, intersection, and complement are useful in deriving a number of compound algebraic statements. Denote by $U(X, Y, Z)$ some algebraic statement constituted by the sets X, Y, and Z and being itself a set. Let $B(X, Y, Z)$ be some another

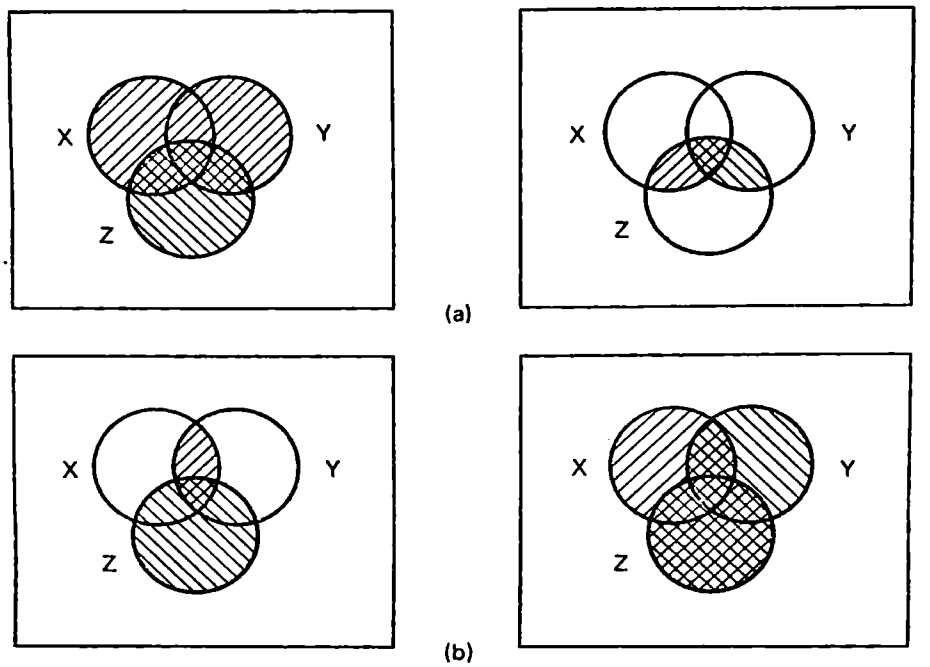


Fig. 2.7. Venn diagrams for the statements (*a*) $(X \cup Y) \cap Z = (X \cap Z) \cup (Y \cap Z)$, and (*b*) $(X \cap Y) \cup Z = (X \cup Z) \cap (Y \cup Z)$

algebraic statement constituted by the same sets. If both statements represent one and the same set, then they may be equated thus yielding some algebraic identity

$$U(X, Y, Z) = B(X, Y, Z) \tag{2.31}$$

Such identities very often come handy in transforming algebra statements. Some of them are provable by Venn diagrams.

(1) Figure 2.7*a* shows two Venn diagrams for the expressions $(X \cup Y) \cap Z$ and $(X \cap Z) \cup (Y \cap Z)$. These diagrams are seen to define the same set, so that we shall be justified to write

$$(X \cup Y) \cap Z = (X \cap Z) \cup (Y \cap Z) \tag{2.32}$$

This statement is analogous to the distributive law of elementary algebra $(a + b)\, c = ac + bc$.

(2) In elementary algebra we cannot substitute multiplication for summation in the distributive law since this would lead to the absurd

expression $(ab) \dashv c = (a \dashv c)\,(b \dashv c)$. The things are quite different in set algebra.

Figure 2.7*b* shows the Venn diagrams for two expressions $(X \cap Y) \cup Z$ and $(X \cup Z) \cap (Y \cup Z)$. Both expressions give one and the same set so that we may write

$$(X \cap Y) \cup Z = (X \cup Z) \cap (Y \cup Z) \tag{2.33}$$

(3) It is an easy matter to prove that if $Y \subseteq X$ then

$$X \cap Y = Y, \quad X \cup Y = X \tag{2.34}$$

Indeed, all elements of Y are at the same time elements in X. Hence the intersection of these sets, i.e. the part common to both X and Y, coincides with Y. The set Y does not contribute to the union $X \cup Y$ any element that would not belong to X. Consequently, $X \cup Y$ coincides with X.

(4) Taking in (2.34) $Y = X$ and observing that $X \subseteq X$, we find

$$X \cap X = X, \quad X \cup X = X \tag{2.35}$$

In some cases Venn diagrams prove inconvenient for visualizing set-algebraic statements. Below we give a more general method. Let as before U (X, Y, Z) and B (X, Y, Z) denote two algebraic expressions obtained with the operations of union, intersection and complement on the sets X, Y, and Z. In order to prove that U = B it suffices to demonstrate that U $\subseteq$ B and B $\subseteq$ U. In turn, to demonstrate that U $\subseteq$ B one needs to verify that from $x \in$ U it follows that $x \in$ B. Likewise, to demonstrate that B $\subseteq$ U one needs to verify that from $x \in$ B it follows that $x \in$ U. We resort to this method to prove two more additional statements.

(5) To prove that

$$\overline{X \cup Y} = \bar{X} \cap \bar{Y} \tag{2.36}$$

we assume that $x \in \overline{X \cup Y}$, that is, $x \notin X \cup Y$. This implies that $x \notin X$ and $x \notin Y$, that is, $x \in \bar{X}$ and $x \in \bar{Y}$. Hence, $x \in \bar{X} \cap \bar{Y}$. Assume now that $y \in \bar{X} \cap \bar{Y}$, i.e. $y \in \bar{X}$ and $y \in \bar{Y}$. This implies that $y \notin X$ and $y \notin Y$, i.e. $y \notin X \cup Y$. Hence, $y \in \overline{X \cup Y}$.

(6) The statement

$$\overline{X \cap Y} = \bar{X} \cup \bar{Y} \tag{2.37}$$

will be proved by reducing both parts of it to the same form. Taking the complement of both parts yields $\overline{\overline{X \cap Y}} = \overline{\bar{X} \cup \bar{Y}}$. The left-hand side of this expression yields $X \cap Y$. This result will be obtained by transforming the right-hand side by (2.36). The two last identities are known as the DeMorgan laws.

2.3 Orderings and the Cartesian Product

2.3.1 Ordered Sets

The *ordered set* is the one in which each element occupies a certain place. Elements of such a set are called accordingly the first element (component or coordinate), the second element, and so on. Examples of ordered sets are people in a queue, and the set of words in a phrase. In such sets, the place of each element is quite definite and cannot be altered at will.

In problems of engineering value, this definiteness is often a matter of agreement. For example a state of a large system is often described by a set of parameters assuming numerical values. Every state of the system is just a set of numbers. To avoid specification at each set, designers agree which parameter comes first, second, and so on, i.e. represent the collection of parameters in the form of an ordered set. By way of example, if we denote by h the altitude of an aircraft flight, and by v its speed, then the ordered pair $x = (h, v)$ will describe the state of this aircraft.

Sometimes the number of elements in an ordered set is referred to as its length, and use is made of parentheses to define such sets. For example, the set

$$a = (a_1, a_2, \ldots, a_n) \tag{2.38}$$

is an ordered set of length n with components $a_1, \ldots, a_n$. Ordered sets of length 2 are called ordered pairs, those of length 3 are called ordered triplets, those of length 4 ordered quadruplets, and so forth. In the general case an ordered set of length n is called an n-tuplet. A set of length one is called a singleton, and the empty ordered set is denoted by (Λ). Unlike ordinary sets, an ordered set may contain identical members, say two identical words in a phrase.

In the following we shall confine ourselves to considering ordered sets whose elements are real numbers. Such ordered sets are often referred to as points of space or vectors. For example, the ordered pair (a_1, a_2) may be viewed as a point in plane or a vector pointing from the origin to the given point, as shown in Fig. 2.8*a*. The components of this pair will be the projections of the vector on the axes *1* and *2*,

$$\mathrm{Pr}_1 (a_1, a_2) = a_1 \quad \text{and} \quad \mathrm{Pr}_2 (a_1, a_2) = a_2$$

An ordered triplet (a_1, a_2, a_3) may be viewed as a point of three-dimensional (3-D) space or a vector in this space drawn from the origin to this point, as shown in Fig. 2.8*b*. The projections of this vector onto the coordinate axes are as follows:

$$\mathrm{Pr}_i (a_1, a_2, a_3) = a_i, \quad i = 1, 2, 3$$

We may also speak of the projection of this vector onto two axes, say *1* and *2*, i.e. onto the coordinate plane. It is not hard to see that this projection is an ordered pair,

$$\mathrm{Pr}_{12}\,(a_1,\ a_2,\ a_3) = (a_1,\ a_2)$$

Generalizing the concepts outlined above, we shall treat an ordered n-member set of real numbers $(a_1, \ldots, a_n)$ as a point in an ima-

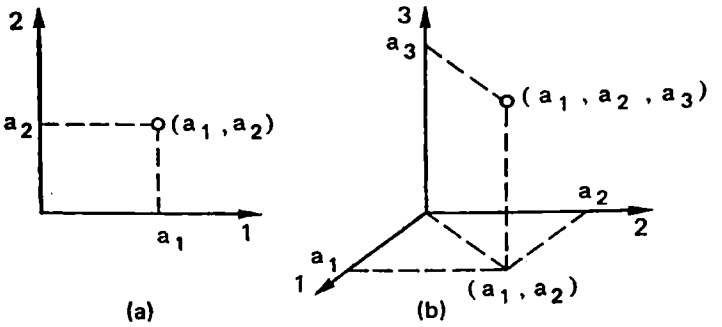


Fig. 2.8. Projections of an ordered pair and a triplet

ginary n-dimensional space, or as an n-vector of *hyperspace*. The components of this n-vector will be defined as projections onto the respective axes,

$$\mathrm{Pr}_i\, a = a_i, \quad i = 1, \ldots, n \tag{2.39}$$

If $i, j, \ldots, l$ are the numbers of axes with $1 \leqslant i < j < \ldots < l \leqslant n$, then the projection of an n-vector **a** onto the axes $i, j, \ldots, l$ will be

$$\mathrm{Pr}_{i, j, \ldots, l}\, \mathbf{a} = (a_i,\ a_j,\ \ldots,\ a_l) \tag{2.40}$$

2.3.2 The Cartesian Product

The Cartesian product (sometimes called a direct product or direct sum) of two sets X and Y is the set, denoted by $X \times Y$, of all pairs (x, y) such that x is an element of X and y is an element of Y. Formally, it is defined as

$$X \times Y = \{(x, y):\ x \in X,\ y \in Y\} \tag{2.41}$$

Example 2.9. Let $X = \{1, 2\}$, $Y = \{1, 3, 4\}$. Then $X \times Y = \{(1, 1), (1, 3), (1, 4), (2, 1), (2, 3), (2, 4)\}$. The geometric representation of this set is given in Fig. 2.9*a*.

Example 2.10. Let X and Y be sections of the real axis. The Cartesian product $X \times Y$ is presented by the hatched rectangle in Fig. 2.9*b*. This example shows that the properties of the Cartesian product differ from those of ordinary product, specifically, the Carte-

sian product changes when the order of multiplicands changes, i.e.

$$X \times Y \neq Y \times X \tag{2.42}$$

The notion of Cartesian product may readily be extended to a larger number of sets. For the sets $X_1, X_2, \ldots, X_r$, this will be $X_1 \times X_2 \times \ldots \times X_r$ consisting of ordered sets of length r such that the first element belongs to X_1, second to X_2, and so on.

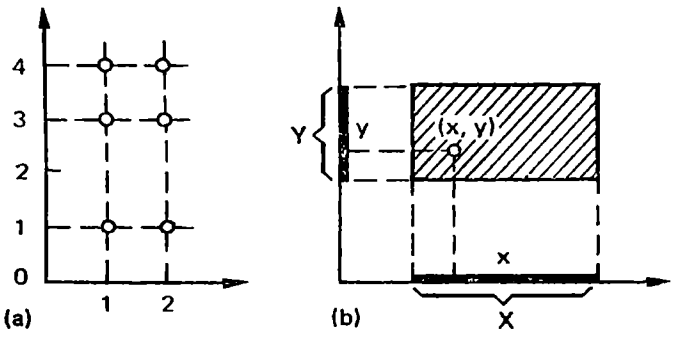


Fig. 2.9. Geometric illustration of Cartesian products

The concept of *power of a set* is a particular case of the Cartesian product. Let M be a set, then the s-th power of M, denoted by M^s, will be the Cartesian product of s identical sets M,

$$M^s = M \times M \times \ldots \times M \tag{2.43}$$

This definition may be extended to any integer non-negative s after letting

$$M^1 = M \quad \text{and} \quad M^0 = \{\Lambda\}_i^1 \tag{2.44}$$

where $\{\Lambda\}$ is an empty ordered set.

If R is the set of real numbers, then the plane of coordinate geometry, in this notation, becomes $R^2 = R \times R$, and the coordinate geometry representation of 3-dimensional space involves $R^3 = R \times R \times R$.

2.3.3 Projection of a Set

Projection is closely connected with the projection of ordered sets and if applied to a set implies that its members are ordered sets of equal length. If M is a set consisting of ordered sets of length s, then the projection of M is the set of projections of its ordered sets.

Example 2.11. Let $M = \{(1, 2, 3, 4, 5), (2, 1, 3, 5, 5), (3, 3, 3, 3, 3), (3, 2, 3, 4, 3)\}$. Then $\mathrm{Pr}_2 M = \{2, 1, 3\}$ and $\mathrm{Pr}_{2,4} M = \{(2, 4), (1, 5), (3, 3)\}$.

It is an easy matter to verify that if $M = X \times Y$, then

$$\mathrm{Pr}_1 M = X \quad \text{and} \quad \mathrm{Pr}_2 M = Y \tag{2.45}$$

and if $Q \subseteq X \times Y$, then

$$\mathrm{Pr}_1 Q \subseteq X \quad \text{and} \quad \mathrm{Pr}_2 Q \subseteq Y \tag{2.46}$$

2.4 Relations

2.4.1 Definition

Consider two sets X and Y. The elements of these sets may be related by producing pairs (x, y). If the method of such a relation is specified, that is for each element x in X an element y in Y is pointed out with which it is associated, then the sets X and Y are said to be related [9, 10]. It is not necessary that all elements of X and Y be involved in the relation.

In order to specify a relation, one needs to point out the set X whose elements are associated with elements of another set, the set Y whose elements are related to elements of the first set, and a set $Q \subseteq X \times Y$ defining a law by which this relation is established, i.e. a method of counting of all pairs (x, y) taking part in the relation.

Thus a relation is a triplet of sets

$$q = (X, Y, Q) \tag{2.47}$$

where $Q \subseteq X \times Y$. The first component in this triplet is called the initial domain of the relation, the second, Y, the co-domain and Q is the graph. The last term will be made clear in discussing the notion of function, a special case of relation.

Each relation has two more sets closely connected to it, namely, the set $\mathrm{Pr}_1 Q$ called the domain of definition including the elements of X involved in the relation and the set $\mathrm{Pr}_2 Q$ called the range of relation and constituted by the elements of Y involved in the relation.

If $(x, y) \in Q$, then we say that element y is related to element x. Geometrically this situation is represented by an arrow from x to y.

Example 2.12. Let $X = \{1, 2\}$, $Y = \{3, 5\}$ so that $X \times Y = \{(1, 3), (1, 5), (2, 3), (2, 5)\}$. This set allows for 16 distinct relations. Some of them follow

$$Q_1 = \{(1, 3)\}; \quad \mathrm{Pr}_1 Q_1 = \{1\}, \quad \mathrm{Pr}_2 Q_1 = \{3\}$$

$$Q_2 = \{(1, 3), (1, 5)\}, \quad \mathrm{Pr}_1 Q_2 = \{1\}, \quad \mathrm{Pr}_2 Q_2 = \{3, 5\} = Y$$

Example 2.13. A small company has two trucks, α and β, operated on a two-shift basis and a bus, γ, for occasional use. The truck β is under repair. The staff includes three drivers, a, b, and c, one of

which, c, has taken a month holiday. An assignment of the drivers to automobiles is a relation that may be as follows:

$$q = (\{a, b, c\}, \{\alpha, \beta, \gamma\}, \{(a, \alpha), (a, \gamma), (b, \alpha)\})$$

Geometrically it is represented in Fig. 2.10*a*. It is defined on a and b, but not defined on c, consequently, the domain is the set $\{a, b\}$ and the range is the set $\{\alpha, \gamma\}$.

2.4.2 Inverse Relation

For each relation $q = (X, Y, Q)$, $Q \subseteq X \times Y$, there exists an inverse relation where elements $x \in X$ related to elements $y \in Y$ are determined. Symbolically, the inverse relation is denoted by

$$q^{-1} = (Y, X, Q^{-1}) \qquad (2.48)$$

with $Q^{-1} \subseteq Y \times X$.

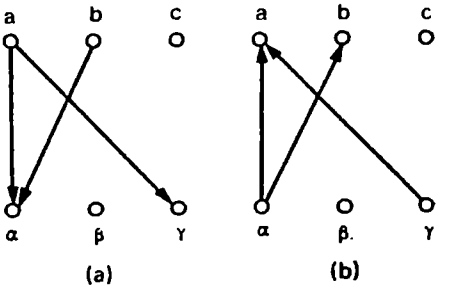


Fig. 2.10. Geometric representation of (*a*) direct and (*b*) inverse relations

Example 2.14. The inverse relation for Example 2.13 will be the assignment of vehicles to the drivers $(\{\alpha, \beta, \gamma\}, \{a, b, c\}, \{(\alpha, a), (\alpha, b), (\gamma, a)\})$, shown in Fig. 2.10*b*.

The geometric representation of an inverse relation is seen to result by reverting the arrows in the geometric representation of the direct relation. The immediate implication is that the inverse relation of an inverse relation is the direct relation,

$$(q^{-1})^{-1} = q \qquad (2.49)$$

2.4.3 Composition of Relations

Consecutively applied relations form a *composition*. In symbolic terms, a composition involves three sets X, Y, and Z on which two relations are defined

$$\left.\begin{aligned} q &= (X, Y, Q), \quad Q \subseteq X \times Y \\ p &= (Y, Z, P), \quad P \subseteq Y \times Z \end{aligned}\right\} \qquad (2.50)$$

where the range of the first relation is the domain for the second, namely,

$$\mathrm{Pr}_2 Q = \mathrm{Pr}_1 P \qquad (2.51)$$

The first relation defines for any $x \in \mathrm{Pr}_1 Q$ some, possibly not unique, element $y \in Y$. By the definition of the composition of relations, for $y \in Y$ we are now to find $z \in Z$ using the second relation. Thus, the composition of relations relates to each x of the

domain $\mathrm{Pr}_1 Q$ one or a few elements z from the range of the second relation $\mathrm{Pr}_2 P$.

We shall denote a composition of relations q and p by $q\,(p)$ and the graph of the composition by $Q \circ P$, that is,

$$q\,(p) = (X,\ Z,\ Q \circ P),\quad Q \odot P \subseteq X \times Z \tag{2.52}$$

Example 2.15. If q is a relation defining an assignment of drivers to vehicles, and p is a relation defining an assignment of the vehicles to certain routes, then the relation $q\,(p)$ defines an assignment of drivers to routes.

Naturally the operation of composition may be extended to more than two relations.

2.5 Mappings and Functions

2.5.1 Mappings

Let X and Y be some sets and $\Gamma \subseteq X \times Y$ with $\mathrm{Pr}_1 \Gamma = X$. The triplet (X, Y, Γ) defines some relation with the property that the domain of definition $\mathrm{Pr}_1 \Gamma$ coincides with the initial domain, i.e. X, and consequently this relation is defined everywhere on X. In other words for every $x \in X$ there exists $y \in Y$ so that $(x, y) \in \Gamma$. This relation defined everywhere on X is called the *mapping* of X onto Y written symbolically as

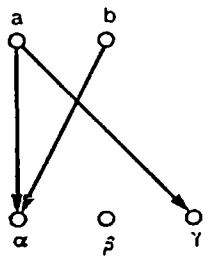


Fig. 2.11. Geometric representation of a mapping

$$\Gamma\colon X \to Y \tag{2.53}$$

The term mapping is often understood as a one-to-one correspondence. We shall not adhere to this rule, but rather shall assume that mapping Γ relates each element $x \in X$ to some subset

$$\Gamma x \subseteq Y \tag{2.54}$$

called the image of element x. A law defining how to map X onto Y is given by the set Γ.

Example 2.16. If we drop driver c from the consideration in Example 2.13, then we obtain a mapping $\Gamma\colon X \to Y$ where $X = \{a,\ b\}$ is the set of drivers, $Y = \{\alpha,\ \beta,\ \gamma\}$ the set of vehicles, and $\Gamma = \{(a,\ \alpha),\ (a,\ \gamma),\ (b,\ \alpha)\}$ an assignment of drivers to vehicles. The geometric representation of this mapping is shown in Fig. 2.11.

Let us now look at some properties of mappings. Assume $A \subset X$. For every $x \in A$, the image of x is the set $\Gamma x = Y$. The collection of all elements of Y that are images Γx for all $x \in A$ we call the

Image of A and denote ΓA. By this definition

$$\Gamma A = \bigcup_{x \in A} \Gamma x \tag{2.55}$$

If A_1 and A_2 are subsets of X, then

$$\Gamma (A_1 \cup A_2) = \Gamma A_1 \cup \Gamma A_2 \tag{2.56}$$

However, the statement

$$\Gamma (A_1 \cap A_2) = \Gamma A_1 \cap \Gamma A_2 \tag{2.57}$$

is valid only if the map Γ: $X \to Y$ is single-valued (one-one). In the general case

$$\Gamma (A_1 \cap A_2) \subseteq \Gamma A_1 \cap \Gamma A_2 \tag{2.58}$$

These relations may be readily generalized over a greater number of subsets A_i. For instance, if $A_1, \ldots, A_n$ are subsets of X, then

$$\Gamma \left(\bigcup_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} \Gamma A_i \tag{2.59}$$

$$\Gamma \left(\bigcap_{i=1}^{n} A_i \right) \subseteq \bigcap_{i=1}^{n} \Gamma A_i \tag{2.60}$$

Since mapping is a particular case of relation, the concepts of inverse mapping and composite mapping apply here as well.

2.5.2 Maps on One Set

An important particular case of mapping is the one when sets X and Y coincide. The mapping Γ: $X \to X$ maps X onto itself and is defined by the pair

$$(X, \Gamma) \tag{2.61}$$

where $\Gamma \subseteq X^2$. A detailed investigation of such mappings can be found in graph theory. Here we only touch upon some operations on such mappings.

Let Γ and Δ map a set X onto X. In accordance with the rule given in Sect. 2.4 the composition of these mappings will be defined as

$$(\Gamma\Delta)\, x = \Gamma (\Delta x) \tag{2.62}$$

In a particular case of $\Delta = \Gamma$ we get the maps

$$\Gamma^2 x = \Gamma (\Gamma x) \tag{2.63}$$

$$\Gamma^3 x = \Gamma (\Gamma^2 x), \quad \text{and so on} \tag{2.64}$$

Thus in the general case, for any $s \geqslant 2$

$$\Gamma^s x = \Gamma (\Gamma^{s-1} x) \tag{2.65}$$

If we denote

$$\Gamma^0 x = x \tag{2.66}$$

then the relation (2.65) may be extended to negative s, namely,

$$\Gamma^0 x = \Gamma(\Gamma^{-1}x) = \Gamma\Gamma^{-1}x = x \tag{2.67}$$

This implies that $\Gamma^{-1}x$ is an inverse mapping, so that

$$\Gamma^{-2}x = \Gamma^{-1}(\Gamma^{-1}x), \text{ and so on} \tag{2.68}$$

Example 2.17. Suppose that X is a set of people. For each $x \in X$, the set of its children will be denoted by Γx. Then $\Gamma^2 x$ is the set of

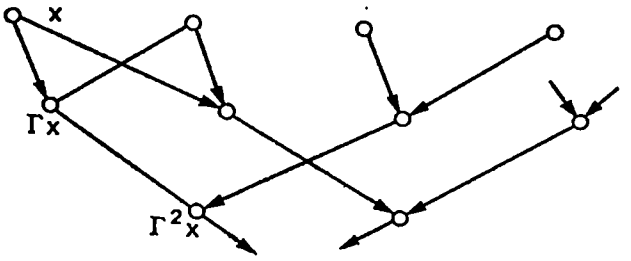


Fig. 2.12. Genealogical tree

grandchildren of x, $\Gamma^3 x$ the set of grandgrandchildren of x, $\Gamma^{-1}x$ the set of parents of x, and so on.

Representing people by points and connecting them by arrow from x to Γx we display the genealogic tree as shown in Fig. 2.12.

Example 2.18. Consider a chess play. Denote by x some position on the chessboard and by X the set of feasible positions. Then for any $x \in X$, Γx implies the set of positions at which one can arrive from x making one move according to the rules. Also, $\Gamma x = \varnothing$ if x is a checkmate or stalemate position, $\Gamma^3 x$ is the set of positions that can be obtained from x in three moves, and $\Gamma^{-1}x$ is the set of positions capable of producing the given position in one move.

For mappings given on one set, some more names will be used in subsequent discussions. For example, if $x \in X$ represent the states of a dynamic system, the mapping Γx will be viewed as a set of states to which the system can be carried from a given state. An appropriate term may be the transformation of states of the dynamic system.

2.5.3 The Notion of Function

A mapping

$$f: X \to Y \tag{2.69}$$

is called a *function* if it is one-one or injective, i.e. for any pairs $(x_1y_1) \in f$ and $(x_2y_2) \in f$ from $x_2 = x_1$ there follows $y_2 = y_1$. Thus, the mapping in Fig. 2.13*a* is a function, whereas that in Fig. 2.13*b* is not.

The definition of a mapping and examples above suggest that the sets X and Y may consist of members of any nature. In problems of system theory and control, however, one-one mappings onto the set of real numbers R are of main interest. The one-one mapping f defined by (2.69) is referred to as a real-valued function if $Y \subseteq R$.

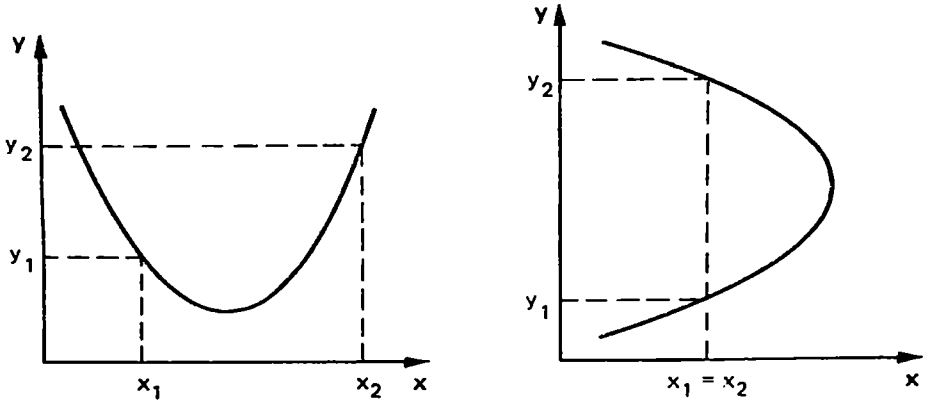


Fig. 2.13. To the concept of function

Separate classes of functions are a subject of study in many a mathematical discipline (algebra, trigonometry, etc.). We consider only the general, most fundamental properties of function without touching upon the properties of special classes.

Example 2.19. To travel from one city to another one can use railway, bus, or airplane at a cost of fare 7, 9, and 12 monetary units, respectively. The fare may be represented as a function of transportation type. Here the domain $X = \{\text{r, b, a}\}$ and the range $Y = \{7, 9, 12\}$.

The function f: $X \to Y$ obtained for this example may be written as $f = \{(\text{r}, 7), (\text{b}, 9), (\text{a}, 12)\}$.

The value y in each pair $(x, y) \in f$ is called the value of function at x under f, written $y = f(x)$. This leads to the following formal definition of a function

$$f = \{(x, y) \in X \times Y: y = f(x)\} \tag{2.70}$$

The symbol f is used therefore in two ways:

(i) to define a set whose elements are the pairs (x, y) taking part in the relation; and

(ii) to denote, in the form $f(x)$, $y \in Y$ related to the given $x \in X$.

The formal definition of a function in the form of the statement (2.70) tells us how to specify a function.

(1) Listing all pairs (x, y) constituting the set f as this has been done in Example 2.19 will suffice. This method is reasonable if X is a finite set. For convenience, the pairs (x, y) are arrayed into a table.

(2) In many cases, both X and Y are sets of real or complex numbers. Then $f(x)$ is understood as a formula, i.e. as an expression containing the mathematical operations (addition, multiplication, taking logarithms, etc.) that should be carried out on $x \in X$ to obtain y.

Example 2.20. Given $X = Y = R$ and $f = \{(x, y) \in R^2;\ y = x^2\}$, then $f(x) = x^2$.

Some situations call for different formulas for distinct subsets of X. Let $A_1, \ldots, A_n$ be pairwise disjoint subsets of X. Denote by $f_i(x)$, $i = 1, \ldots, n$, the formula defining y at $x \in A_i$. Then

$$\begin{aligned} f(x) &= f_1(x), \quad \text{at } x \in A_1 \\ &\cdots\cdots\cdots \\ &= f_n(x), \quad \text{at } x \in A_n \end{aligned} \tag{2.71}$$

For instance, the function $y = f(x) = |x|$ may be given as

$$\begin{aligned} y &= x, \quad \text{for } x \geqslant 0 \\ &= -x, \quad \text{for } x < 0 \end{aligned}$$

(3) If X and Y are sets of real numbers, then the elements $(x, y) \in f$ can be represented as points in the plane R^2. The full set of such points makes up the *graph* of $f(x)$.

If we let $X = U \times V$ in (2.69), we obtain a function of two variables u and v, written $f(u, v)$, where $u \in U$ and $v \in V$. Formally the definition of this function is as follows:

$$f = \{(u, v, y)\} \in U \times V \times Y : y = f(u, v)\} \tag{2.72}$$

Functions of three or more variables may be defined in a similar manner.

One more concept, known as the *restriction* of a function, is intimately connected with the concept of function. Suppose that $f\colon X \to Y$ is an arbitrary function and A is an arbitrary set. The restriction of f to A is a function f_A containing only those pairs $(x, y) \in f$ for which $x \in A$, and consequently $(x, y) \in A \times Y$. Written symbolically, the restriction is

$$f_A = f \cap (A \times Y) \tag{2.73}$$

The operation of restriction is often used to represent functions with infinite domain of definition X in table form. The set A is usually a sample of equidistant values x of X. The respective restriction f_A may readily be arranged into a table. Available tables of logarithms and trigonometric functions are constructed in this way.

2.5.4 The Inverse Function

The concept of an inverse function applies to a mapping f: $X \to Y$ which is single-valued, i.e. for any $(x_1, y_1) \in f$ and $(x_2, y_2) \in f$ if $x_2 = x_1$, then $y_2 = y_1$, and is one-one, i.e. if $x_2 \neq x_1$, then $y_2 \neq y_1$. When these conditions are satisfied, the mapping f: $X \to Y$ is onto and one-one, i.e. defines a function $y = f(x)$. The inverse mapping f^{-1}: $Y \to X$ is also one-one and defines a function $x = f^{-1}(y)$ called the *inverse* for $y = f(x)$.

Figure 2.14*a* shows a function having an inverse, whereas Fig. 2.14*b* shows a function that has not. When a function f is specified ana-

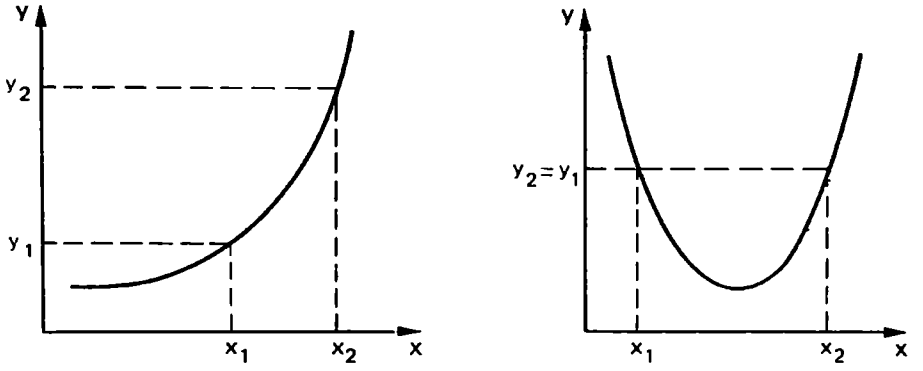


Fig. 2.14. To the notion of universe function

lytically, the argument of both direct and inverse functions is commonly denoted by x. Therefore to find the inverse one has to resolve the equation $y = f(x)$ for x and replace x with y and y with x, so that the inverse becomes $y = f^{-1}(x)$.

2.5.5 Functions of Time

The concept of a function of time is based on a set $T \subseteq R$ whose elements t are called instants of time. A characteristic feature of time is associated with its directivity. This implies that if $t_1, t_2 \in T$ and $t_1 < t_2$, then t_1 precedes t_2. In other words, T is an ordered set.

The function of time defines a mapping f of a set of time instants T onto the set of real numbers R, viz.,

$$f: T \to R \tag{2.74}$$

The elements of f are pairs (t, x) also denoted as $x(t)$, for which $t \in T$ and $x \in R$. Each pair of this mapping gives the value of the function at instant t and is referred to as an *event* or an *instantaneous value* of the function. The full set of pairs (t, x), i.e. values $x(t)$, for all $t \in T$ represents the function of time. A further refinement of

a function of time is associated with the refinement of its domain of definition, i.e. the form of set T.

If $T = R$, i.e. t can take on any real value from $-\infty$ to ∞, then the function $x(t)$ is a *continuous function of time*. An example may be the sine function of time $x(t) = A \sin(\omega t + \varphi)$ describing voltage in an a.c. circuit.

As a rule, distant time instants in the future and in the past are not of much interest. Therefore investigators commonly restrict $x(t)$ to a finite interval $t_1 < t \leqslant t_2$, and denote this semiclosed interval as $(t_1, t_2]$. Semiclosed time intervals are convenient in that

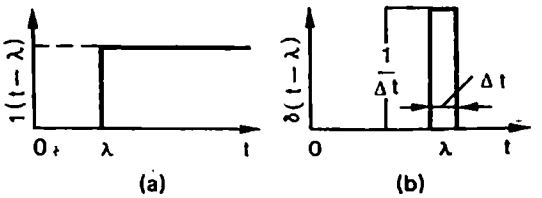


Fig. 2.15. A unit step and an impulse function

they allow for consecutive chain linking with one another. For example, if we divide the interval $(t_1, t_2]$ by a time instant t' into two intervals $(t_1, t']$ and $(t', t_2]$, then there is no doubt which interval t' belongs to.

The restriction of a function $x(t)$ given on the interval $-\infty < t < +\infty$ to the interval $(t_1, t_2]$ is sometimes called a *segment of function* $x(t)$ and denoted by $x_{(t_1, t_2]}$, viz.,

$$x_{(t_1, t_2]} = \{x(t) : t \in (t_1, t_2]\} \tag{2.75}$$

The operation of contraction is often performed with the aid of a special time function, known as the *unit step*,

$$\begin{aligned} 1(t - \lambda) &= 0, \text{ for } t \leqslant \lambda \\ &= 1, \text{ for } t > \lambda \end{aligned} \tag{2.76}$$

shown in Fig. 2.15*a*. For instance, the voltage applied to a device connected to the mains at time $t = \lambda$ will be

$$u(t) = 1(t - \lambda)\, x(t) = 1(t - \lambda)\, A \sin(\omega t + \varphi)$$

Another function of time in wide use is the impulse function $\delta(t - \lambda)$ defined as

$$\begin{aligned} \delta(t - \lambda) &= 0, \quad \text{at } t \neq \lambda \\ &= \infty, \text{ at } t = \lambda \end{aligned} \tag{2.77}$$

$$\int_{\lambda - \varepsilon}^{\lambda + \varepsilon} \delta(t - \lambda)\, dt = 1, \quad \varepsilon > 0 \tag{2.78}$$

This function may be viewed as the limiting case for the rectangular pulse of width Δt and height $1/\Delta t$ (Fig. 2.15*b*) that appears at instant $t = \lambda$ as $\Delta t \to 0$.

A useful approximation of the impulse function is the Gaussian function (discussed in Sect. 5.3)

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\lambda)^2}{2\sigma^2}\right]$$

which becomes an impulse function $\delta(t)$ as $\sigma \to 0$.

The impulse function allows one to evaluate instantaneous values of $x(t)$ at fixed time instants. For example, if $t_1 < \lambda < t_2$, then

$$\int_{t_1}^{t_2} x(t)\,\delta(t-\lambda)\,dt = x(\lambda)\int_{\lambda-\varepsilon}^{\lambda+\varepsilon} \delta(t-\lambda)\,dt = x(\lambda) \qquad (2.79)$$

If the set T is a collection of natural numbers

$$\ldots, -2, -1, 0, 1, 2, \ldots, n, \ldots$$

then we say of a *discrete time function*. The elements of T are usually labeled by n, so that a pair (n, x), also written $x[n]$ or x_n, gives the

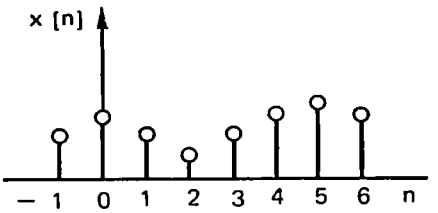


Fig. 2.16. Discrete time function

value of the function at a time instant n. An example of a discrete function of time is shown in Fig. 2.16.

2.5.6 The Notion of a Functional

When we spoke of the mapping f: $X \to Y$ as of a real-valued function, we did not impose any constraints on the character of elements in X. In simple problems, the members of both sets X and Y are real values. Each pair $(x, y) \in f$ relates one real number x to another real number y. A situation of practical value is when X is a set of functions, and Y is a set of real values. This case leads to the concept of a functional which will be exposed by way of an example.

Figure 2.17 shows a line $y = f(x)$ connecting two fixed points A and B. For a point moving from A to B the time in motion depends on the profile of this line, i.e. on the type of function $f(x)$. If we denote by $F(x)$ the set of functions representing the line AB and

by T the set of real numbers t defining the time in motion, then the dependence of motion time on the type of function may be written as the mapping

$$J\colon\ F\ (x) \to T \tag{2.80}$$

The elements of set J will be pairs $(f\ (x),\ t)$ with $f\ (x) \in F\ (x)$ and $t \in T$. This is the set-theoretic approach to saying that the real value t represents the functional J of function $f\ (x) \in F\ (x)$, written as

$$t = J\ [f\ (x)] \tag{2.81}$$

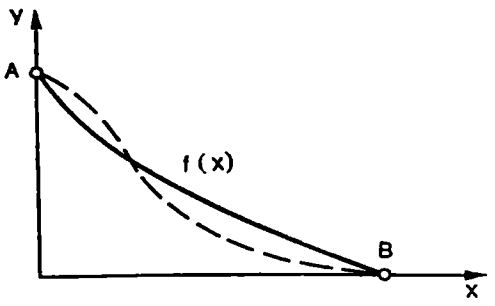


Fig. 2.17. Fastest descent line

In problems of control, functionals find their use as performance criteria. For example, in a system of Fig. 2.17, time en route may be treated as a performance measure of a given function $f\ (x)$. More often than not such problems seek for an *optimal control* as such that brings the respective performance criterion to an extremum. In our case, optimization of the type of function $f\ (x)$ boils down to satisfying the condition

$$\min_{f \in F} J\ [f\ (x)] \tag{2.82}$$

that is, to seeking for the shortest time. In the calculus of variations, the problem of finding the equation of the path along which a particle will fall from one point to another in the shortest time is known as the brachistochrone problem. The correct solution is a cycloid through the two points.

2.5.7 The Notion of an Operator

The *operator* L is a mapping

$$L\colon\ X \to Y \tag{2.83}$$

in which X and Y are sets of functions with elements $x\ (t)$ and $y\ (t)$, so that the members of set L are pairs $(x\ (t),\ y\ (t))$. The operator L is said to transform a function $x\ (t)$ into another function

$$y\ (t) = L\ [x\ (t)] \tag{2.84}$$

By way of example we take the differentiation operator p which is a rule that tells us how to assign to a function $f\ (x)$ another function $f'\ (x) = \mathrm{d}f(x)/\mathrm{d}x$, written

$$f'\ (x) = p\ [f\ (x)]$$

In control systems, the role of an operator is often performed by the controlled system itself, which transforms an input signal $x(t)$

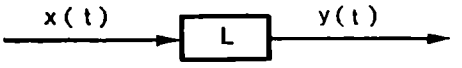


Fig. 2.18. Representation of a controlled system by an operator

into output signal $y(t)$ according to some law L, as shown schematically in Fig. 2.18.

2.6 Relations Revisited

2.6.1 The Formal Definition of a Relation

We recall that a relation on a set X is a subset $X \times X$ and a relation between sets X and Y is a subset of $X \times Y$. It is clear that a relation between X and Y may be treated as a relation on the union $X \cup Y$, therefore we consider only relations on a single set. In symbolic terms this definition may be disclosed as follows.

Let a map (X, R) be a relation. For an element $y \in Rx$ we say that y is related to x under R and write

$$y \, R \, x \tag{2.85}$$

In Example 2.17 R means "to be children of this person".

Remark. Using (2.61) for a mapping given on a single set we obtain that the relation is a pair of sets (X, R) with $R \subseteq X^2$. The elements of X^2 are ordered pairs, therefore we are justified by saying that the relation is a set of ordered pairs. Since each pair relates only two elements of X^2, this relation may be termed *binary*.

More generally a relation can be referred to as a pair of the set (X, R) with $R \subseteq X^n$. The elements of set X^n are ordered n-tuplets, which allows us to call this relation n-ary. Specifically, a set of ordered triplets may be referred to as a *ternary relation*. In what follows we confine ourselves to binary relations, calling them relations.

Consider six fundamental properties of relations, assuming that x, y, and z are elements in X. A relation is

(1) reflexive: $x \, R \, x$ is true;
(2) antireflexive: $x \, R \, x$ is false;
(3) symmetric: $x \, R \, y \rightarrow y \, R \, x$;
(4) antisymmetric: $x \, R \, y$ and $y \, R \, x \rightarrow x = y$;
(5) non-symmetric: if $x \, R \, y$ is true, then $y \, R \, x$ is false;
(6) transitive: $x \, R \, y$ and $y \, R \, z \rightarrow x \, R \, z$.

In the ensuing sections we look at some important types of relations.

2.6.2 Equivalence Relations

Elements of a set may be viewed as equivalent if under some consideration one of them may be replaced with another. An equivalence relation is then said to exist on the set of these elements. Examples of equivalent relations are "being on the same course" given on the set of students of a department, "having identical remainder if divided by 3" on the set of natural numbers, "being parallel" on the set of coplanar straight lines, and "being congruent" on the set of triangles.

An equivalence relation R on a set X is a relation which satisfies the following conditions, for all $x, y, z \in X$:

each element is equivalent to itself;

a statement that two elements are equivalent does not require pointing out which element comes first;

two elements equivalent to a third are equivalent to each other.

If we take the symbol "$\equiv$" to denote equivalence, then these three conditions reduce to the following symbolic statements:

(1) reflexivity, $x \equiv x$, or $x R x$;

(2) symmetry, $x \equiv y \rightarrow y \equiv x$, or whenever $x R y$ then also $y R x$;

(3) transitivity, $x \equiv y$ and $y \equiv z \rightarrow x \equiv z$, or if $x R x$ and $x R y$, then $x R z$.

Thus, a relation R is an equivalence relation if it satisfies the RTS conditions, i.e., if it is reflexive, symmetric, and transitive.

The equivalence relation is closely connected with the partition of a set considered in Sect. 2.2. Suppose that X is a set on which an equivalence relation is defined. For instance, let X be a set of students in a course, and the equivalence relation be "being in one group". A subset of elements equivalent to some element $x \in X$ will be called an equivalence class. In our illustrative example, the group where student Jones studies is equivalent to student Jones.

Let J be a set of indices. Denote by $\{A_j \subseteq X: j \in J\}$ a set of equivalence classes for a set X. Clearly all elements of one equivalence class are equivalent to one another (transitivity) and any element $x \in X$ can appear in only one class. Then X is the union of disjoint sets A_j, so that the total system of classes $\{A_j \subseteq X: j \in J\}$ is a partition of X. Thus, any equivalence relation on a set X induces some partitioning of the set into classes A_j.

We say that an equivalence relation on a set X is induced by the partitioning, if for any x and y of X the relation $x \equiv y$ holds if and only if both x and y belong to the same class A_j of this partitioning. We invite the reader to look again through examples of Sect. 2.2 and compare them with the examples of this section for clear perception of the above concepts.

In this book we shall use the symbol "$\equiv$" for equivalence, although sometimes an arbitrary equivalence relation is denoted "$\sim$". Other special equivalence symbols are "$=$" for equality, $\|$ for parallelism between lines, and $\rightleftarrows$ for logic equivalence.

2.6.3 Order Relations

Relations defining some order of elements in a set are frequent situations. For example, we distinct the concepts 'earlier' and 'later' in situations where elements of the set are states of a dynamic system. We also differentiate between the concepts 'greater than' and 'less than' and use the symbols $>$ or $<$ to denote them if these elements are numbers. Likewise the symbols $\subseteq$ or $\subset$ are used to indicate the order of things for sets and subsets.

In all these cases the elements of a set X or a group of elements can be arranged to a certain order, in other words, an *order relation*, or *ordering*, can be introduced on X.

Two order relations are distinguished, the total order and the partial order. Using the symbols $\leqslant$ and $<$, we say that a partial order possesses the following properties:

$x \leqslant x$ (reflexivity),

$x \leqslant y$ and $y \leqslant x \rightarrow x = y$ (antisymmetry),

$x \leqslant y$ and $y \leqslant z \rightarrow x \leqslant z$ (transitivity);

and a total order possesses the properties as follows:

$x < x$ does not hold (antireflexivity),

$x < y$ and $y < x$ are mutually exclusive (nonsymmetry),

$x < y$ and $y < z \rightarrow x < z$ (transitivity).

A set X is an ordered set if any two elements of this set x and y are comparable, i.e. obey the *law of trichotomy*

$$x = y \quad \text{or} \quad x < y \quad \text{or} \quad y < x$$

2.6.4 Relations of Dominance

For $x, y \in X$ (where X is often a set of people) we shall say that x dominates y, written $x \gg y$, if x is superior to y in some aspect. For example, x may be a sportsman or a team that defeats a sportsman or a team y, or a property that we prefer to that denoted by y.

We shall say that a relation of dominance exists on X if for every two elements of this set the following conditions are satisfied:

no element dominates itself, i.e. $x \gg x$ is false,

in a pair, one element dominates the other, i.e. $x \gg y$ and $y \gg x$ are mutually exclusive statements.

A dominance relation is not transitive. Indeed, if team x defeats team y, and y defeats z, then team x will not necessarily defeat z.

EXERCISES

2.1. Define the set $Y \setminus X$ of Example 2.2.

2.2. Indicate by section lines the set $Y \setminus X$ of Example 2.3.

2.3. Let R be the set of real numbers and

$$X = \{x \in R: 0 \leqslant x \leqslant 1\} \quad \text{and} \quad Y = \{y \in R: 0 \leqslant y \leqslant 2\}$$

What are the sets $X \cup Y$, $X \cap Y$, and $X \setminus Y$?

2.4. Draw Venn diagrams for the sets

$$A = \{(x, y) \in R^2: x^2 + y^2 \leqslant 1\}$$
$$B = \{(x, y) \in R^2: x^2 + (y - 1)^2 \leqslant 1\}$$

Which geometric figures represent the sets $A \cup B$, $A \cap B$, and $R^2 \setminus A$?

2.5. Use (2.34) to prove that $X \cap \varnothing = \varnothing$, $X \cup \varnothing = X$, $U \cap X = X$, and $U \cup X = U$.

2.6. Show the sets $X \times Y$ and $Y \times X$ of Exercise 2.3 in the plane $R^2 = R \times R$.

2.7. Give geometric interpretation of the sets $A \times R$ and $R \times A$, for $A = [2, 3]$.

2.8. What geometric figure represents the intersection of the set of all rectangles and the set of all rhombuses?

2.9. Find $\mathrm{Pr}_1 M$ and $\mathrm{Pr}_2 M$ for the set $M = \{(x, y) \in R^2: (x - 2)^2 + y^2 = 1\}$.

2.10. Let $U = \{x_1, x_2, x_3\}$ be a universe of discourse with the proper subsets $X = \{x_1, x_2\}$; $Y = \{x_2, x_3\}$, and $Z = \{x_3\}$. Define by enumeration the following sets: $X \times X$, $Z \times Z$, $X \times Y$, $Y \times X$, $X \times Y \cap Y \times X$, and $X \times Y \cup Y \times X$.

2.11. Given the subsets of R^2, $X = \{(x, y) : x \geqslant 0\}$, $Y = \{(x, y): y \geqslant 0\}$, and $Z = \{(x, y): x + y \geqslant 1\}$. Plot Venn diagrams for X, Y, Z, $X \cup Y$, $\overline{X \cup Y}$, $\overline{X \cap Y}$, $X \cap Y$, $X \cap Z$, $\overline{X \cap Z}$, $X \cap Y \cap Z$, and $X \cap Y \cap \overline{Z}$.

2.12. Determine the set $X \times Y$ for the subsets of R^2 given as follows:

$$X = \{(x, y) : 2x + y = 1\}, \; Y = \{(x, y): x - y = 0\}.$$

2.13. List all 16 relations in Example 2.12. Determine $\mathrm{Pr}_1 Q$ and $\mathrm{Pr}_2 Q$ for each of them.

2.14. Give the inverse f^{-1} for f in Example 2.20.

2.15. Determine the composition of f and g given in R^2 as $f = \{(x, y) : y = x^2\}$, $g = \{(y, z) : z = \sin y\}$.

3

FUNDAMENTALS OF GRAPH THEORY

3.1 Basic Concepts and Definitions

3.1.1 Set-Theoretical Definition of a Graph

Pictorially a graph can be represented by a diagram in which there is a finite set of dots called *vertices* and line segments connecting all or some of these dots, called *edges* or *arcs* [11,12]. Mathematically, a graph

$$G = (X, U) \tag{3.1}$$

consists of two sets: a finite set X of vertices, and a finite set U of edges. Each edge is identified with a pair of vertices. If the edges of a graph G are identified with ordered pairs of vertices, then G is called a *directed* or an *oriented* graph. Otherwise G is called an undirected on a nonoriented graph.

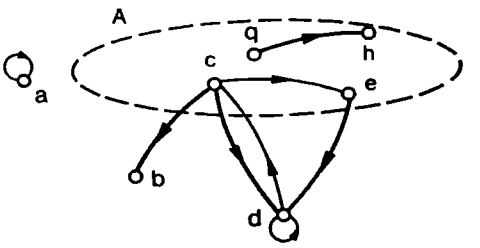


Fig. 3.1. An oriented graph

Figure 3.1 shows a directed graph in which $X = (a, b, c, d, e, g, h)$ and the directed edges (often called arcs) connect the following end vertices (initial vertex, terminal vertex pairs): $((a, a), (c, b), (c, d), (c, e), (d, c), (d, d), (e, d), (g, h)$. Examples of directed graphs are genealogical relationships (see Fig. 2.12) and dominance relations of members of a tournament. For undirected graphs, these may be a road map and an electrical circuit.

Sometimes it is more convenient to define a directed graph by observing that the set U of directed edges connecting the elements of the set X maps this set into itself. Therefore, we may assume a graph to be specified if there is the set X of its vertices and a method Γ of mapping X onto X. Thus, the graph G is the pair (X, Γ) that consists of a set X and a mapping on this set,

$$G = (X, \Gamma) \tag{3.2}$$

With reference to the graph represented in Fig. 3.1 this mapping is defined as follows:

$$\Gamma a = a, \quad \Gamma b = \varnothing, \quad \Gamma c = \{b, d, e\}, \quad \Gamma d = \{d, c\},$$
$$\Gamma e = d, \quad \Gamma g = h, \quad \Gamma h = \varnothing$$

As can be seen this definition of a graph coincides completely with the definition of a relation on a set.

A subgraph G_A of $G = (X, \Gamma)$ is a graph whose set of vertices A is a subset of X, and the set of edges is a subset of the set of edges of G. By way of mapping

$$G_A = (A, \Gamma_A) \tag{3.3}$$

where

$$A \subseteq X, \text{ and } \Gamma_A x = (\Gamma x) \cap A \tag{3.4}$$

If all the vertices of a graph G are present in a subgraph G_A, then G_Δ is called a *spanning subgraph* or *partial graph* of G. Symbolically

$$G_\Delta = (X, \Delta) \tag{3.5}$$

where

$$\Delta x \subseteq \Gamma x \tag{3.6}$$

The spanning graph for the graph shown in Fig. 3.1 is emphasized by heavy edges.

Example 3.1. If $G = (X, \Gamma)$ is the map of Soviet motor ways, then the map of highways in the European part of the Soviet Union is a subgraph of this graph, and the map of all highways represents a spanning subgraph.

A walk in a graph G is a sequence of edges $\mu = (u_1, \ldots, u_k)$ such that the terminal vertex of one edge is the initial vertex of the next edge. Alternatively, a walk can be considered as a finite sequence of vertices $a, b, \ldots, m$, and denoted by $\mu = (a, b, \ldots, m)$, or called an a-m walk with a and m referred to as end or terminal vertices of this walk.

A walk is a *trail* if all its vertices are distinct. A trail is *open* if all its end vertices are distinct, otherwise it is *closed*. An open trail is a *path* if all its vertices are distinct.

The number of edges in a path is called the *length of the path*, $l(\mu) = k$. Sometimes a number $l(u_i)$ called the *length of the edge* u_i is assigned to each edge u_i. Then the length of the path is

$$l(\mu) = \sum_{u \in \mu} l(u) \tag{3.7}$$

A closed trail is a *circuit* if all its vertices except end vertices are distinct.

All edges having the same pair of end vertices are called parallel edges. The end vertices of an edge need not be distinct. If $u_k = (x_i, x_i)$, then the edge u_k is called a *self-loop* at vertex x_i. A graph is called a *simple graph* if it has no parallel edges or self-loops. By way of example, in the graph of Fig. 3.1, the sequence (e, d, c, b) is a walk, the sequence (c, e, d, c) is a circuit, and the sequence (d, d) is a self-loop.

An edge is said to be *incident* on its end vertices. Two vertices are *adjacent* if they are the end vertices of some edge. An edge is said to be incident *out of* its initial vertex and incident *into* its terminal vertex. A vertex is called *isolated* if no edge is incident on it.

Let $G = (X, U)$ be a directed graph with no parallel edges. Let $X = (x_1, x_2, \ldots, x_n)$. The *adjacency matrix* $\mathbf{R} = [r_{ij}]$ of G is an $n \times n$ matrix with $[r_{ij}]$ defined as follows:

$$r_{ij} = \begin{cases} 1, & \text{if } (x_i, x_j) \in U \\ 0, & \text{otherwise} \end{cases}$$

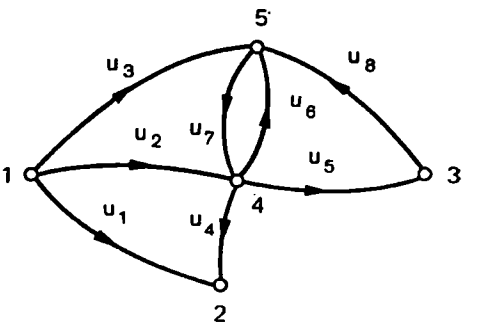


Fig. 3.2. Graph without self-loops

Consider a graph G with n vertices and m edges and having no self-loops. The *all-vertex incidence matrix* $\mathbf{S} = [s_{ij}]$ of G has n rows, one for each vertex, and m columns, one for each edge. The element s_{ij} of S is defined as follows:

$$\begin{aligned} s_{ij} &= 1, \quad \text{if } u_j \text{ is incident out of } x_i \\ &= -1, \text{ if } u_j \text{ is incident into } x_i \\ &= 0, \text{ if } u_j \text{ is not incident on } x_i \end{aligned}$$

If self-loops are present in the graph, then the all-vertex incidence matrix should be partitioned in two semimatrices, positive and negative.

Given below are the adjacency matrix and the all-vertex incidence matrix of the graph shown in Fig. 3.2.

Adjacency matrix

$j=$	1	2	3	4	5
$i=1$	0	1	0	1	1
2	0	0	0	0	0
3	0	0	0	0	1
4	0	1	1	0	1
5	0	0	0	1	0

All-vertex incidence matrix

$j=$	1	2	3	4	5	6	7	8
$i=1$	1	1	1	0	0	0	0	0
2	−1	0	0	−1	0	0	0	0
3	0	0	0	0	−1	0	0	1
4	0	−1	0	1	1	1	−1	0
5	0	0	−1	0	0	−1	1	−1

3.1.2 Nonoriented Graphs

If the edges of a graph are nonoriented, i.e. the pairs of vertices with which each edge is identified are not ordered, the graph is called *undirected* or *nonoriented*. The concepts of walk, path, and circuit are defined as for the oriented graph above.

Figure 3.3*a* shows a nonoriented graph whose vertices are labelled by numerals, and the edges by letters.

A nonoriented graph G may be described by specifying its sets of vertices and edges as a pair (X, U). However, a more convenient way is to describe such a graph by means of its adjacency matrix or all-

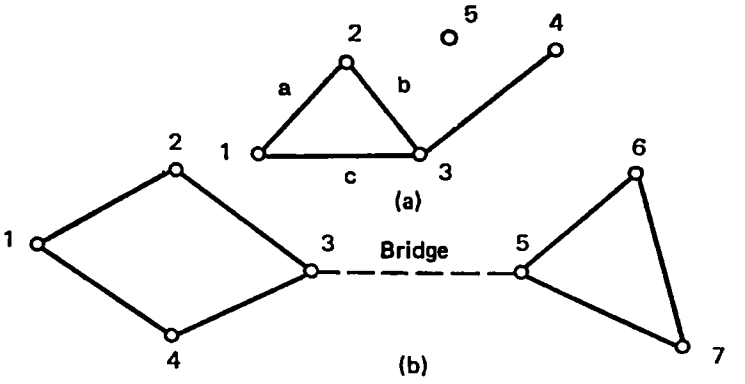


Fig. 3.3. Undirected graphs

vertex incidence matrix. These matrices are constructed similar to how they were derived for oriented graphs with the difference that the edges incident on a vertex are not distinct.

Two vertices are *adjacent* if they are the end vertices of some edge. If two edges have a common end vertex, then these edges are said to be *adjacent*.

For the graph in Fig. 3.3*a*, the adjacency and all-vertex incidence matrices have the form

Adjacency matrix

$$\begin{array}{c} j= \\ i=1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

All-vertex incidence matrix

$$\begin{array}{c} X= \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad U$$

The number of edges incident on a vertex x_i is called the *degree of the vertex*, and it is denoted by $\deg(x_i)$, d_x, or $d(x_i)$. Sometimes the degree of a vertex is also called its *valency*. With reference to the graph in Fig. 3.3*a*, we have $d_1 = 2$, $d_2 = 2$, $d_3 = 3$, $d_4 = 1$, and $d_5 = 0$. A vertex of degree *1* is called a *pendant vertex* (vertex *4*). The only edge incident on a pendant vertex is called a *pendant edge*. A vertex of degree 0 is called an *isolated vertex* (vertex *5*). By definition, a self-loop at a vertex x contributes 2 to the degree of x.

Theorem 3.1. The sum of the degrees of the vertices of a graph G is equal to $2m$, where m is the number of edges of G.

Proof. Since each edge is incident on two vertices, it contributes two units to the sum of the degrees of the graph G. Consequently, all the edges together contribute $2m$ to the sum of the degrees of G.

Corollary. The number of vertices of odd degree in any graph is even.

The concepts of subgraph and spanning subgraph of a nonoriented graph are similar to that of the oriented graph.

An important concept of graph theory is that of *connectedness*. Two vertices x_i and x_j are said to be *connected* in a graph G if there exists an $x_i = x_j$ path in G.

If a graph is disconnected, then its vertex set can be partitioned into subsets such that the vertex induced subgraphs are connected, and no vertex in subset X_i is connected to any vertex in subset X_j, $i \neq j$. Such subgraphs are called the components of G. By way of illustration, if we exclude isolated vertex *5* from the graph in Fig. 3.3*a*, then the resulting graph will be connected. The graph in Fig. 3.3*b* is disconnected and has two components. This graph can be converted into a connected graph by bridging vertices *3* and *5* with an edge. Eliminating the bridge renders the connected graph a disconnected one.

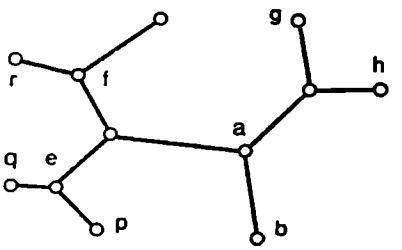


Fig. 3.4. A tree

Orientation of edges is not taken into account in determining the connectedness of an oriented graph. The graph shown in Fig. 3.1 is disconnected. However, its subgraph constituted by the vertices b, c, d, and e is connected.

In a directed graph G, a vertex x is *strongly connected* to a vertex y if in G there exists a directed path from x to y and a directed path from y to x. If x is strongly connected to y, then, obviously, y is strongly connected to x.

A graph is said to be *acyclic* if it has no circuits. A *tree* is a *connected acyclic graph*, as shown in Fig. 3.4. A tree of a graph G is a connected acyclic subgraph of G.

For a set of vertices a, b, c, . . . we can construct a tree as follows. Assume one of the vertices, say a, to be the initial vertex, referred to as the *root* of a tree. Join the root with the nearest vertices, say b, c, and d. Connect these vertices to the vertices, adjacent to them, say e, f, g, and h, and so on. Therefore, a tree can by constructed by adding edges at pendant vertices. This leads us to establish a relationship between the number of vertices and the number of edges in a tree.

A simple tree consists of two vertices connected by an edge. Each time we add one more edge to a tree we also add one more vertex. Consequently, a tree with n vertices has $n - 1$ edges.

3.1.3 Isomorphic Graphs

One and the same graph may have a few geometrical representations, as those shown in Fig. 3.5*a*. These graphs are called *isomorphic*. More rigorously, two graphs G_1 and G_2 are said to be isomorphic if there exists a one-to-one correspondence between their vertex sets and a one-to-one correspondence between their edge sets so that the corresponding edges of G_1 and G_2 are incident on the corresponding vertices of G_1 and G_2.

If there are two graphs, how can we know whether they are isomorphic or not? This question is rather hard to answer. The practical

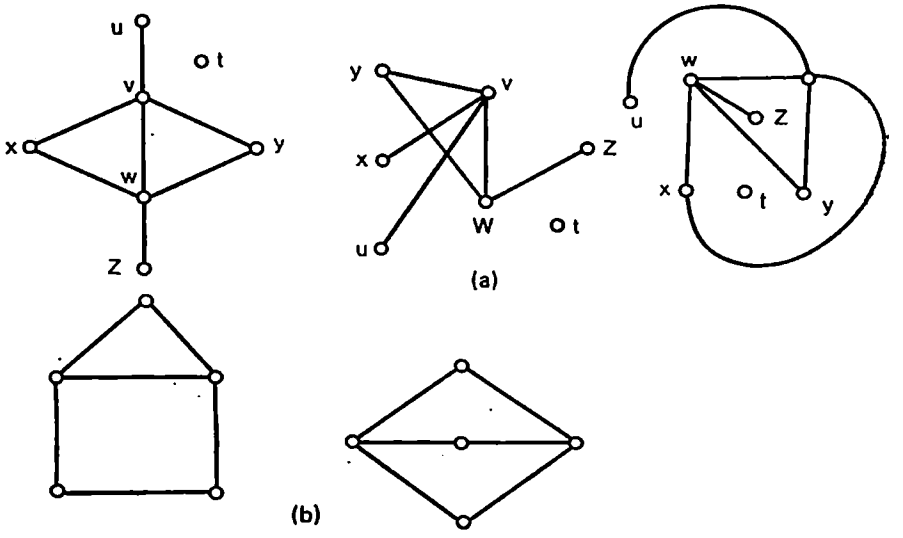


Fig. 3.5. (*a*) Isomorphic and (*b*) non-isomorphic graphs

approach usually involves a preliminary evaluation of some parameters for both graphs. These may be the number of vertices, number of edges, number of components, and the sequence of degrees of the vertices in descending order. If some of these parameters of the two graphs differ, then these graphs are distinct. However, if all the parameters of both graphs coincide, this does not guarantee that these graphs are isomorphic. By way of illustration, Fig. 3.5*b* shows two graphs with all the parameters coincident, which nevertheless are distinct.

3.1.4 Graphs and Relations

A binary relation on a set $X = \{x_1, x_2, \ldots\}$ is a collection of ordered pairs of elements of X. A most convenient way of representing a binary relation R on a set X is by a directed graph, the vertices of which stand for the elements of X and the edges stand for the ordered pairs of elements of X defining the relation R. Assume that $R \equiv$ an

order relation, i.e. which is introduced on a graph $G = (X, \Gamma)$ by the statement that any two vertices x and y satisfying the condition $x \leqslant y$ are connected by a path from x into y. We say that vertex x precedes y or that y follows x. This definition reflects on a graph all the properties of the order relation.

R is reflexive. The statement that the condition

$$x \leqslant x \text{ is true} \tag{3.8}$$

implies that the vertex is equivalent to itself, i.e. the condition $x \equiv x$. This condition may be viewed as a self-loop at x (Fig. 3.6*a*).

R is transitive. The statement

$$x \leqslant y,\ y \leqslant z \rightarrow x \leqslant z \tag{3.9}$$

implies that vertices x, y and z occur successfully along the same path (Fig. 3.6*b*).

R is antisymmetric. We demonstrate that the condition

$$x \leqslant y,\ y \leqslant x \rightarrow x \equiv y \tag{3.10}$$

is true. The left-hand side of this expression implies that there exist a path from x to y and a path from y to x. However, this means that the graph has a closed trail involving vertices x and y, as shown in Fig. 3.6*c*.

From the right-hand side of the condition (3.10) it follows that the vertices lying in one closed trail are equivalent. We shall view this

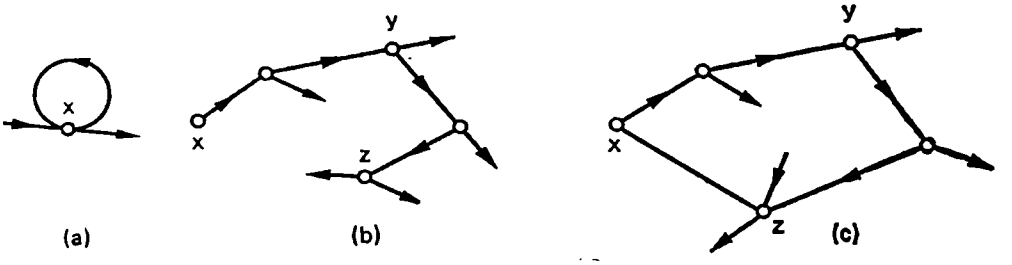


Fig. 3.6. Illustrating the order relations

observation as a definition of equivalence on a graph and demonstrate that this definition satisfies all the three conditions of an equivalence relation. The conditions of reflexivity $x \equiv x$ and symmetry $x \equiv y \rightarrow y \equiv x$ are obvious and follow from the above definition of equivalence. The condition of transitivity $x \equiv y$, $y \equiv z \rightarrow x \equiv z$ is also obvious, since if a graph has a circuit with vertices x and y, and a circuit with vertices y and z, then it has a circuit containing vertices x and z (Fig. 3.6*c*).

Thus, an order relation and an equivalence relation together define a certain graph.

A relation of strict order may be also introduced on a graph. In this case, for any two vertices $x < y$ there exists a path from x into

y. The condition of transitivity $x < y < z \rightarrow x < z$ implies as above that vertices x, y, and z occur successfully along the same path. The condition of antireflexivity ($x < x$ is false) implies that the graph has no self-loops, and the condition of non-symmetry ($x < y$, $y < x$ are mutually exclusive) implies that the graph has no circuits. Thus a relation of strict order defines a graph having no circuits.

3.1.5 Characteristics of Graphs

The graph theoretical solution of many engineering problems reduces to evaluation of characteristics of the respective graphs. We shall not dwell on the engineering applications of graph theory in this volume, however, the knowledge of the principal characteristics of graphs may be of value in other disciplines, and we shall touch on them below.

Nullity. Consider a graph G with m edges, n vertices and r components. The nullity of G, denoted by $\nu(G)$, is defined as $m - n + r$. This number has an interesting physical meaning, namely, it indicates the maximum number of independent circuits in a graph. Specifically, it may be used to determine the number of independent circuits in electrical circuit design.

Chromatic number. A p-vertex coloring of a graph is the assignment of p distinct colors to the vertices of the graph. A vertex coloring is *proper* if in the coloring no two adjacent vertices receive the same color. The chromatic number of a graph, denoted by $\gamma(G)$, is the minimum number p for which G is p-vertex colorable. G is p-chromatic if $\gamma(G) = p$.

If $\gamma(G) = 2$, the graph is called bichromatic. A necessary and sufficient condition that a graph be bichromatic is that it has no circuits of odd length. The chromatic number plays an important role in solving problems of efficient use of memory space. However, its evaluation, except for bichromatic graphs, is a hard task often requiring the help of a computer.

Independence (stability) number. Consider a graph $G = (X, \Gamma)$. A subset S of X is an independent set of G if no two vertices of S are adjacent in G. An independent set is also called a stable set.

A stable set S of G is maximum if G has no independent set S' with $|S'| > |S|$. The number of vertices in a maximum stable set of G is called the independence number (stability number) of G. The maximum stable set plays an important role in communication theory.

Vertex covering number. A subset T of X is a vertex cover of G if every edge of G has at least one end vertex in T. If we consider a vertex as covering all the edges incident on it, then a vertex cover of G is a subset of X which covers all the edges of G. Mathematically, for any $x \notin T$, we have $\Gamma x \cap T \neq \varnothing$.

A vertex cover T of G is minimum if G has no vertex cover T' with $|K'| < |K|$. The number of vetices in a minimum vertex cover is called the vertex covering number of G.

3.2 The Shortest Path Problem

The problem of shortest path between two vertices of a connected undirected graph is of great significance for applications relating to operations research and computer science. Such a problem arises in choosing a shortest, fastest, or less expensive route from an available map, or in selecting a most economical method of carrying the dynamic system from one state to another. Of all methods developed for solving such problems the graph-theoretical methods often prove to be easiest.

3.2.1 Problem Statement

Let G is a connected undirected graph $G = (X, U)$ in which each edge is associated with a real positive number called the length of the edge and denoted by $l(u)$. The length of a path in G is the sum of the lengths of the edges in the path. (In particular problems, this may be the length between two end vertices, the fare of moving through this path, time taken in traveling between these end vertices, and the like.)

It is required to find the shortest path μ_{ab} between vertices a and b of G, that is the path having the minimum length.

The length of a shortest a-b path is called the distance from a to b, sometimes denoted by $d(a, b)$.

Before we evaluate the general method of solution for this problem, we consider a solving algorithm for a particular type of problem in which the length of all edges is unity.

3.2.2 Shortest Paths in a Unit-edge Graph

Some applications may be represented by graphs with edges of the same length, taken for convenience equal to unity. Vertices of such a graph are usually states of some system in which all transitions in one step are equivalent. We consider a problem that reduces to evaluation of such a graph below. This problem may provide an illustration of construction of graphs for various specific situations.

Example 3.2 (*Hanoi tower problem*). The first of three vertical rods has m discs threaded on it so that the disc diameter decreases from the bottom disc upwards. Displacing the discs one by one to the remaining rods so that a larger disc should never occur above a smaller one on either rod, thread the discs on the third rod in the origi-

nal order. An additional condition may be to find a solution that requires the least number of steps.

We assign to the discs the numbers in descending order of their diameters, namely, m, $m-1$, . . ., 1. Denote by X, Y and Z the sets of discs on the respective rods at any step. Clearly it suffices to give the sets X and Z only, as Y results as a complement of X and Z to the total number of discs. Elements of set X or Z are disc combinations as follows: 0, 1, 2, 21, 3, 31, 32, 321, 4, 41, 42, 421, 43, 431, 432, 4321, and so on. These combinations can be represented by points

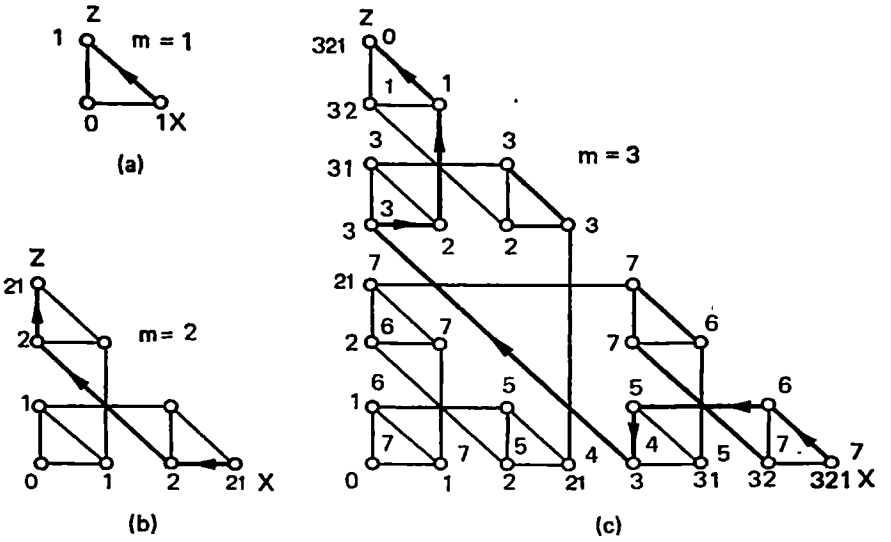


Fig. 3.7. Transition graphs in the Hanoi tower problem

of the X-Z plane as shown in Fig. 3.7. Connecting these points by edges indicating the possible displacements of discs at each step, we arrive at a graph in which we can trace a path, including a shortest path, to move from the initial point of the graph to the terminal point.

The construction of this graph is carried out by considering a transition from m discs to $m + 1$ discs. At $m = 1$ the feasible states are given by the set (1, 0), (0, 0), (0, 1); the corresponding graph is depicted in Fig. 3.7*a*. The shortest path from initial to terminal state (1, 0)-(0, 1) is indicated by an arrow.

To derive the general rule we assume that the graph has been constructed for m discs; we call this graph an m-graph. It is not hard to verify that this graph will be in the form of a triangle, shown in Fig. 3.8*a*, where the internal links are not clear yet. What will be the graph for $m + 1$ discs?

We recall that disc $m + 1$ may occupy only the bottom position on either rod. The remaining m discs may be moved arbitrarily in accordance with the m-graph leaving the $(m + 1)$th disc intact. Con-

sequently, the $m + 1$-graph (Fig. 3.8*b*) will consist of three m-graphs for the three rods (encircled) at which disc $m + 1$ may be placed. It remains to find the transitions from one m-graph to another that correspond to movements of disc $m + 1$.

The $(m + 1)$th disc may be placed only on a free rod, i.e. only when all m other discs are on the other rod. Accordingly, Fig. 3.8*b* shows the feasible transitions for disc $m + 1$, denoted by encircled numbers of the rod holding m smaller discs. This diagram illustrates the

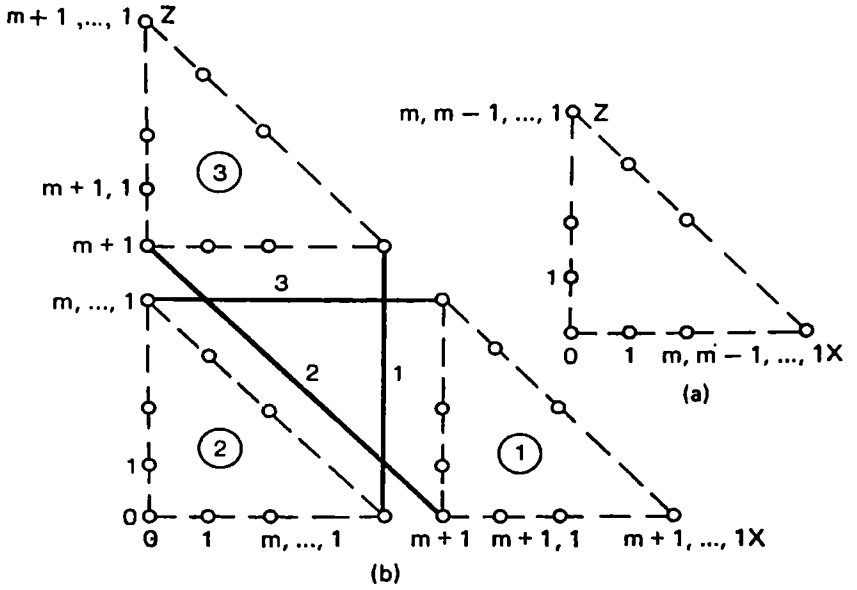


Fig. 3.8. Transition from an m-graph to an $(m + 1)$-graph

general principle of transition from m-graph to $m + 1$ graph. The transition graphs for two and three discs are shown in Fig. 3.7*b*,*c*.

Let us consider the relevant shortest path problem. Because the considered graphs are comparatively simple, the shortest path can be readily found by trying out all feasible paths. However, for complex graphs a systematic approach should be developed.

We assign to each vertex the length of the shortest path λ_i from this vertex x_i to the terminal vertex. For a unit-edge graph the assignment may be as follows:

(1) assign 0 to the terminal vertex x_0,

(2) assign 1 to the adjacent vertices,

(3) assign $\lambda_i + 1$ to all vertices adjacent to the vertices distant λ_i from the terminal vertex. This process continues until the initial vertex is labeled. This procedure results in the shortest path at the initial vertex. The route of this path can be traced by moving from the initial vertex in decreasing index order.

An example of length assignment and shortest path evaluation for $m = 3$ is given in Fig. 3.7*c*.

It is worth noting that this method of evaluation of the shortest path is a particular case of seeking an optimal solution by the dynamic programming method. Therefore, it would be instructive to revisit this example after studying Sect. 8.2 devoted to dynamic programming.

3.2.3 Shortest Path in Arbitrary Edge Graphs

The problem of assigning numerical values to vertices of a graph complicates when edges are arbitrary in length. Here a path through a smaller number of vertices may have a greater length than paths

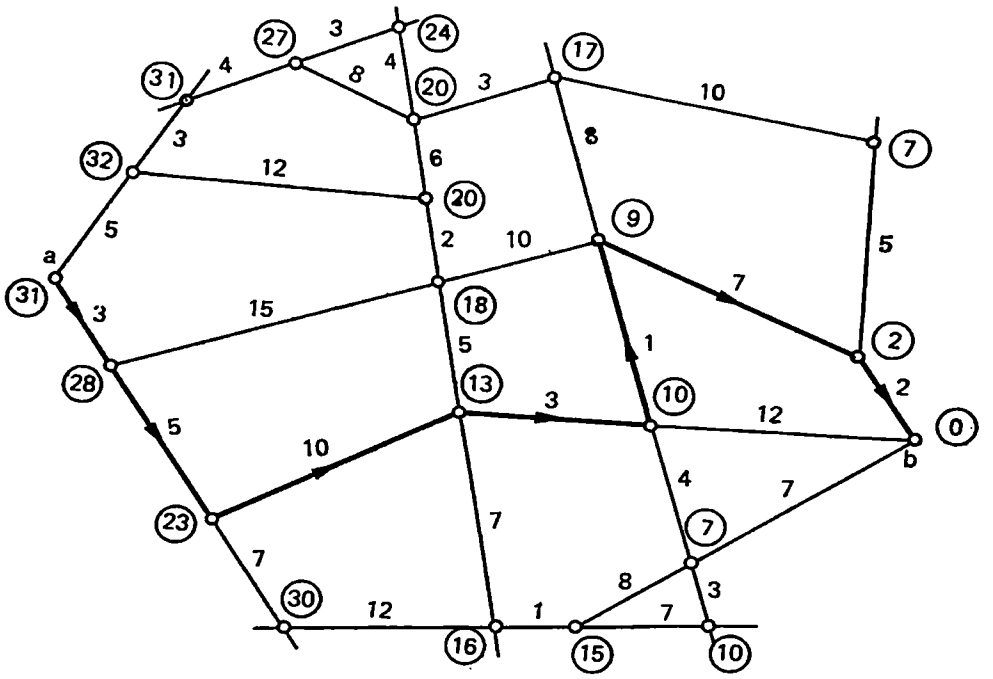


Fig. 3.9. Road map

involving greater number of vertices. For example, the road map represented by a graph in Fig. 3.9 illustrates how a direct path from the vertex marked with an asterisk into the terminal vertex has a length of 12 units, whereas a path through the vertex marked by a triangle is only 10 units long.

A procedure of assignment for this type of graphs is as follows [11,13].

(1) Each vertex is labeled by λ_i. Set the initial vertex label to 0, $\lambda_0 = 0$, and mark it permanent. At the initial step we attach to the remaining vertices $\lambda_i = \infty$ $(i \neq 0)$ and mark them temporary.

(2) Update the labels as follows. Seek for an edge (x_i, x_j) for which $\lambda_j - \lambda_i > l(x_i, x_j)$ and substitute the index $\lambda'_j = \lambda_i + l(x_i, x_j) <$

λ_j for λ_j. Continue the updating of indices until there remains at least one edge for which λ_j can be reduced.

We note two properties of the new labels assigned to the vertices.

i. For any edge (x_k, x_s) in the graph, $\lambda_s - \lambda_k \leqslant l(x_k, x_s)$, because if this is not the case, then λ_s should be reduced.

ii. Let x_p be an arbitrary vertex in the graph. In the above assignment procedure, label λ_p monotonously decreases. Let x_q be the last vertex involved in the reduction. Then $\lambda_p = \lambda_q + l(x_q, x_p)$. Consequently, for an arbitrary vertex x_p with label λ_p there exists a vertex x_q adjacent to x_p and such that $\lambda_p - \lambda_q = l(x_q, x_p)$.

These properties lead us to the following rule for evaluation of the shortest path.

Let $x_n = a$ be the initial vertex with label λ_n. Seek a vertex x_{p_1} such that $\lambda_n - \lambda_{p_1} = l(x_{p_1}, x_n)$. Then seek a vertex x_{p_2} such that $\lambda_{p_1} - \lambda_{p_2} = l(x_{p_2}, x_{p_1})$, and so on until the terminal vertex $x_{p_{k+1}} = x_0 = b$ is achieved. The path $\mu_0 = (x_n, x_{p_1}, \ldots, x_{p_k}, x_0)$ of length λ_n is the shortest path.

To prove this statement we consider an arbitrary path from a to b; $\mu = (x_n, x_{k_1}, \ldots, x_{k_s}, x_0)$ of length $l(\mu)$. According to the label assignment rule, we must have for this path

$$\begin{aligned}
&\lambda_n - \lambda_{k_1} \leqslant l(x_n, x_{k_1}) \\
&\lambda_{k_1} - \lambda_{k_2} \leqslant l(x_{k_1}, x_{k_2}) \\
&\cdots\cdots\cdots\cdots\cdots \\
&\lambda_{k_s} - 0 \leqslant l(x_{k_s}, x_0)
\end{aligned} \tag{3.11}$$

Adding these inequalities termwise yields for any path μ

$$\lambda_n - 0 \leqslant l(\mu) \tag{3.12}$$

The path μ_0 is seen to be the shortest as it satisfies the condition $\lambda_n = l(\mu_0)$.

We illustrated the procedure of shortest path evaluation with reference to Fig. 3.9 showing a road map. The numbers next to the edges indicate time taken to pass each of these roads. The labels at the vertices show the time taken to travel from this vertex to the terminal vertex.

3.2.4 Minimum-Weighted Path Length Graphs

The route optimization problem discussed below is of great practical significance. Assume that there are a few cities that should be connected by a network of highways. For each pair of cities (x, y) the construction cost $l(x, y)$ of the connecting highway is given. The problem is to construct the cheapest possible network. In place of highways the graph can represent communication lines, pipelines, power supply lines, etc. The weight associated with each edge (x, y),

i.e. its cost, is here referred to as the length of this edge, $l(x, y)$. Accordingly our problem may be referred to as a minimum-weighted length graph.

If the set of vertices consists of only three elements, a, b, and c, then it will be sufficient to construct one of the connecting paths, abc, acb, bac. We should eliminate the longest edge, say bc, and construct the path bac.

A graph of minimum length is always a tree, since if it involved a circuit, one of the edges in this circuit could be eliminated without having to disturb the connectedness of all the vertices. Consequently, $n - 1$ edges should be constructed to connect n vertices.

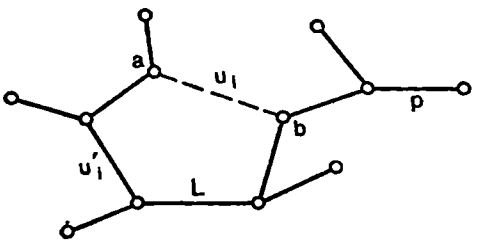


Fig. 3.10. To illustrate the construction of a minimum-length tree

We intend to demonstrate that a graph of minimum length can be constructed by an algorithm described by Ore [12]. First of all connect two vertices with the shortest edge u_1. At each subsequent step add the shortest of edges u_i which does not lead to a circuit. If there is a choice of a few edges of the same length, pick any of them. We shall refer to every tree Q constructed in this way as an *economic tree*. Its length is equal to the sum of the lengths of individual edges

$$l(Q) = l(u_1) + \ldots + l(u_{n-1}) \tag{3.13}$$

Let us demonstrate that no other tree connecting the same vertices can have a length smaller than that of the economic tree Q. Suppose that P is the minimum length tree connecting the vertices under investigation, and Q is an economic tree. Assume also that the edges $u_1, u_2, \ldots, u_{n-1}$ are labeled in the order of their addition into Q, i.e. that they satisfy the condition $l(u_k) \leqslant l(u_{k+1})$. If P does not coincide with Q, then Q has at least one edge that does not belong to P. Let $u_i = (a, b)$ be a first such edge and let $L(a, b)$ be a path in P connecting a and b, as shown in Fig. 3.10. If we add u_i to P, we obtain a circuit and since Q has none, this circuit must involve at least one edge that does not belong to Q. Let it be u_i'. Deleating this edge yields tree P' with the same number of vertices as in P and the length

$$l(P') = l(P) + l(u_i) - l(u_i') \tag{3.14}$$

Since P has the smallest length, then

$$l(u_i) \geqslant l(u_i') \tag{3.15}$$

By construction u_i was a shortest-length edge which when added to the edges $u_1, u_2, \ldots, u_{i-1}$ did not result in a circuit. Since the addi-

tion of u_i' to these edges does not produce a circuit, it follows that

$$l(u_i) = l(u_i') \tag{3.16}$$

consequently, P' like P has the smallest length. However, P' has one extra edge common with Q as compared to P. By repeating this operation several times we arrive at a tree of shortest length that coincides with Q. Hence, Q is the tree of shortest length.

3.3 Flows in a Transport Network

3.3.1 Basic Concepts and Definitions

A *transport network* is a connected directed graph which has no self-loops and which satisfies the following conditions.

1. There is only one vertex with zero in-degree (zero number of edges entering it) s such that $\Gamma^{-1} s = \varnothing$. This vertex is called the *source*.

2. There is only one vertex t with zero out-degree (zero number of edges leaving it) such that $\Gamma t = \varnothing$. This vertex is called the *sink*.

3. Each directed edge u is associated with a non-negative real number $c(u)$ called the *capacity* of the edge.

A transport network is a useful model for the transportation of a commodity from source to sink through communication routes. The capacity of an edge may be considered as representing the maximum rate at which the commodity can be transported along this edge.

A flow in a transport network is an assignment of a nonnegative real number $f(u)$ to each edge. The value $f(u)$ of the flow in edge u may be considered as the rate at which material is transported along u under the flow f subject to the following conditions

$$0 \leqslant f(u) \leqslant c(u) \text{ for every edge } u \text{ in } U \tag{3.17}$$

$$\sum_{u \in U_x^-} f(u) - \sum_{u \in U_x^+} f(u) = 0, \quad x \neq s, \quad x \neq t \tag{3.18}$$

where U_x^- and U_x^+ stand for the sets of edges entering (*in-degree*) and leaving (*out-degree*) a vertex x. Condition (3.17), called the *capacity constraint*, requires that the rate of flow along an edge should not exceed the capacity of the edge. Condition (3.18), called the *conservation condition*, requires that for vertex x, except the source s and sink t, the rate at which material is transported into x is equal to the rate at which it is transported out of x. This implies that the total amount of material transported out of the source is equal to the total amount transported into the sink, namely,

$$\sum_{u \in U_s^+} f(u) = \sum_{u \in U_t^-} f(u) = f_t = \text{val}(f) \tag{3.19}$$

This quantity is called the *value of a flow* f.

Figure 3.11 gives an example of a transport network. The numbers in breaks of edges indicate the capacity of every edge. The arrows indicate the direction of flow and the numbers next to them represent the value of flow. Many problems of logistics systems, power or gas supply networks are formulated and solved as network flow problems, i.e. as problems of maximum flow from s to t.

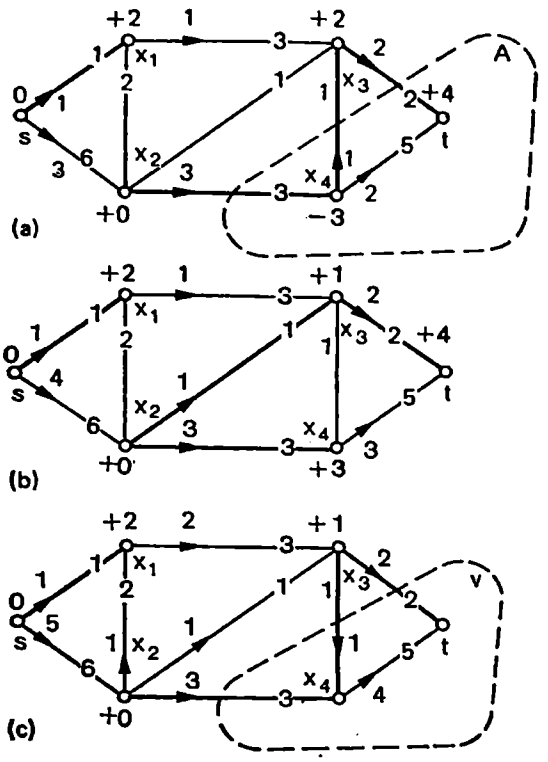


Fig. 3.11. Flow patterns in a transport network

Before presenting an algorithm for finding maximum flows in graphs we introduce the concept of a cut. This is a set of edges whose removal from the graph disconnects vertex s from vertex t, i.e. destroys all the paths leading from s to t. In set-theoretic terms let $A \subset X$ be a set of vertices satisfying the conditions

$$s \notin A, \quad t \in A \tag{3.20}$$

Denote by U_A^- and U_A^+ the sets of vertices incident into and incident out of A respectively. The set $U_A = U_A^- \cup U_A^+$ is then called a *cut*. An example of a cut is given in Fig. 3.11*a*.

Because every parcel of material flowing from s to t has to pass along some edge of the cut, the total flow through the cut must be equal to the flow of this network. For any cut A,

$$f_t = \sum_{u \in U_A^-} f(u) - \sum_{u \in U_A^+} f(u) \tag{3.21}$$

The capacity of a cut A is the sum of capacities of the edges incident into this cut,

$$c(A) = \sum_{u \in U_A^-} c(u) \tag{3.22}$$

Since for any edge $f(u) \leqslant c(u)$, then from (3.21) and (3.22) it follows that

$$f_t \leqslant c(A) \tag{3.23}$$

3.3.2 Maximum Flow Problem

The problem of the maximum flow from s to t in a network is formulated as follows. Given a transport network configuration and capacities of edges in the network, find the maximum flow that can be passed through this network and the distribution of the flow over the edges of the network.

Lemma. For any network N, the maximum flow from s to t is equal to the minimum capacity of a cut separating s from t. The cut is known as the minimum cut.

Proof. Denote by V the cut with the minimum capacity (Fig. 3.11*c*)

$$c(V) = \min_{A} c(A) \tag{3.24}$$

We observed here that for any cut A the flow f_t must satisfy the constraint (3.23). Since f_t is the same for any cut of the network, this flow can be increased until it reaches the value of $c(V)$. Consequently, the value

$$f_t = c(V) \tag{3.25}$$

defines the maximum flow of the transport network.

For all its utility this lemma fails to provide us with a practical procedure for evaluation of the maximum flow. Before we begin discussing such a method we introduce a few more concepts.

An edge u is called *saturated* if $f(u) = c(u)$. A flow f_t is full if every path from s to t contains at least one saturated edge.

The full flow for a given network is not a definite quantity and depends on the direction of flow in the individual edges. For example, Fig. 3.11*a* and *c* gives two distinct distributions of flow in the network. The saturated edges are emphasized by heavy arrows. Both situations represent full flows though of a different value.

We now discuss an algorithm due to Ford and Fulkerson [13] for determining the maximum flow in a transport network. It consists in a gradual increase of flow f_t until it becomes maximum. We assume that the capacities of the edges $c(u)$ are expressed by integer numbers, so that flows through edges will also be integers. The evaluation of the maximum flow is in two phases.

1. *Evaluation of a full flow.* Let $f(u)$ be some distribution of flow over the edges of the network. We seek a path μ from s to t whose edges are all unsaturated and assume

$$\begin{aligned} f'(u) &= f(u) + 1 \qquad \text{for } u \text{ in } \mu \\ &= f(u) \qquad\quad \text{for } u \text{ not in } \mu \end{aligned} \tag{3.26}$$

This changes the flow f_t to a value of $f'_t = f_t + 1 > f_t$. In this way we gradually augment f_t until it becomes a full flow.

Example 3.3. Find the full flow in the transport network of Fig. 3.11*a*. We shall consider one by one paths with unsaturated edges (and mark saturated edges by heavy lines)

$\mu_1 = (s, x_1, x_3, t)$, $f(\mu_1) = 1$, edge (s, x_1) is saturated;

$\mu_2 = (s, x_2, x_4, x_3, t)$, $f(\mu_2) = 1$, edges (x_4, x_3) and (x_3, t) are saturated;

$\mu_3 = (s, x_2, x_4, t)$, to saturate this path we can take $f(\mu_3) = 2$, then edge (x_2, x_4) is saturated.

Now, no path remains from s to t which would contain an unsaturated edge. Consequently, the full flow is

$$f_t = f(\mu_1) + f(\mu_2) + f(\mu_3) = 4$$

2. *Evaluation of the maximum flow.* Let f_t be a full flow, $f(x, y)$ being a flow in arc $u = (x, y)$ from x to y. We shall increase f_t by giving to vertices of the graph labels that indicate a path through which a larger flow can be sent. We assume that all the vertices in the network are numbered.

We start by labeling s with 0. If x_i is a vertex labeled on the previous step, then we attach labels $+i$ to all unlabeled vertices adjacent to x_i and joint with it by unsaturated edges, i.e. vertices y for which

$$(x_i, y) \in U \text{ and } f(x_i, y) < c(x_i, y) \tag{3.27}$$

To all initial vertices y of edges whose final vertex is x_i that are unlabeled and for which

$$(y, x_i) \in U, f(y, x_i) > 0 \tag{3.28}$$

we attach the label $-i$.

This process is repeated until t is labeled, which implies that there exists an s-t path μ with distinct vertices labeled (accurate to the sign) by the numbers of the preceding vertices. If this is the case, then we increase the flow through the edges of this path by one unit in the direction from s to t, that is, we change the flow along edge u in μ for

$$f'(u) = f(u) + 1$$

if the label on x_i is of the form $+j$, and for

$$f'(u) = f(u) - 1$$

if the label on x_i is of the form $-j$.

This process results in a new flow pattern $f_t' = f_t + 1$ with a greater flow. This process should be repeated until this is no longer possible.

If some flow f_t^0 cannot be increased by the above method any longer, i.e. t cannot be labeled, then the existing flow is the maximum flow.

To prove, let V be a set of unlabeled vertices including t. Consequently, V is a cut such that has no edge leaving this set (otherwise some vertices of this cut were labeled by negative labels), and all edges incident on this cut are saturated, namely,

$$\begin{aligned} U_V^- &= U_V,\ U_V^+ = \varnothing \\ f(u) &= c(u) \text{ for } u \in U_V^- \end{aligned} \tag{3.29}$$

$$f_t^0 = \sum_{u \in U_V^-} f(u) - \sum_{u \in U_V^+} f(u) = \sum_{u \in U_V^-} c(u) - 0 = c(V) \tag{3.30}$$

In view of the lemma proved earlier, f_t^0 is the maximum flow and V is the minimum cut.

Example 3.4. A feasible labeling of the vertices of the graph in Fig. 3.11*a* resulted in a label of $+4$ on t. The decreasing sequence of labels $+4$, -3, $+2$, $+0$ defines the path $\mu = (s, x_2, x_3, x_4, t)$ in which the s-t flow should be increased by one unit. This results in a flow pattern shown in Fig. 3.11*b*. Repeating the process of labeling on this diagram, we arrive at the path $\mu = (s, x_2, x_1, x_3, x_4, t)$ in which the flow also should be increased by one unit. The resultant flow pattern is presented in Fig. 3.11*c*. Vertex labeling in this diagram leaves unlabeled only the vertices x_4 and t. Consequently, this flow pattern allows the largest flow f_t^0 from s to t, and the set $V = \{x_4, t\}$ defines the minimum capacity cut. The value of the flow is found as the capacity of V

$$f_t^0 = c(x_2, x_4) + c(x_3, x_4) + c(x_3, t) = 3 + 1 + 2 = 6$$

The maximum flow of a network can be increased by raising the capacity of some of the edges incident into the cut V.

3.3.3 Transportation Problem

The problem in which the cost of transporting material along the edges of a transport network should be minimized is known as the transportation problem. Because a transport network is often a road map for transporting some goods, the solution of this problem yields an optimal transportation pattern, that is an arrangement of routes that assures a minimum cost of transportation or shipments in shortest time. Accordingly the optimization criterion may be cost or time of transportation.

Denote for convenience by $c_{ij} = c(x_i, x_j)$ the capacity of edge (x_i, x_j), and by $d_{ij} = d(x_i, x_j)$ the cost of transporting a unit flow along edge (x_i, x_j).

In terms of graph theory a cost-optimal transportation problem may be formulated as follows.

Given a transport network with the maximum flow f_t^0 and the flow $f_t \leqslant f_t^0$ that should be passed through this network. It is required to find such a distribution of flow over the edges of the network that would minimize the cost of transportation subject to the constraint for every edge $f(x_i, x_j) \leqslant c_{ij}$. Clearly the cost of sending a flow $f(x_i, x_j)$ along edge (x_i, x_j) is $d_{ij}f(x_i, x_j)$.

To solve this problem we shall consider the values d_{ij} as lengths of the respective edges. In this case the cost of sending a flow f along a path μ from s to t will be the product of the length of this path and the flow f, so that the cost minimization problem reduces to the evaluation of the shortest path from s to t in the graph. If no constraints are imposed on the capacity of the edges, then the shortest path is the one providing the minimum cost of the flow.

In the presence of constraints imposed on the capacity the problem is solved in several steps by evaluating partial flows at each step. The general approach to the solution may be as follows.

In a graph $G_1 = (x, \Gamma)$ representing a network with edges $d_{ij} = l(x_i, x_j)$ we seek the shortest path μ_1 from s to t. Suppose c_1 is the capacity of path μ_1. This path accommodates the flow

$$\begin{aligned} f_1 &= f_t \quad \text{if } f_t \leqslant c_1 \\ &= c_1 \quad \text{if } f_t > c_1 \end{aligned} \tag{3.31}$$

When $f_t \leqslant c_1$, the problem is solved and path μ_1 is that of minimum cost for flow f_t.

When $f_t > c_1$, we treat the flow f_1 as a partial flow and move on to the graph G_2 that is obtained from G_1 by replacing its capacities c_{ij} by c'_{ij} as follows:

$$\begin{aligned} c'_{ij} &= c_{ij} - c_1 \quad \text{for } u \text{ in } \mu_1 \\ &= c_{ij} \quad \text{for } u \text{ not in } \mu_1 \end{aligned} \tag{3.32}$$

and eliminating the edges whose new capacity $c'_{ij} = 0$. The flow to be distributed over the graph G_2 is

$$f'_t = f_t - f_1 . \tag{3.33}$$

Now we have the shortest path problem for the flow f'_t in the graph G_2. The solution of this problem yields a path μ_2 of capacity c_2 allowing a partial flow

$$\begin{aligned} f_2 &= f'_t, \quad \text{if } f'_t \leqslant c_2 \\ &= c_2, \quad \text{if } f'_t > c_2 \end{aligned} \tag{3.34}$$

When $f'_t \leqslant c_2$, the problem is solved and the minimum cost flow pattern in G_1 will be passing the flow f_1 along μ_1 and the flow f_2 along μ_2.

If $f'_t > c_2$, then one more graph G_3 should be constructed as before and a new partial flow f_3 should be evaluated. This process is repea-

ted until the sum of partial flows equals f_t. The resultant partial flows sent along the edges of G_1 provide the minimum cost flow pattern for the flow f_t.

To illustrate this algorithm we consider a minimum cost transportation problem in which there are m sources of supply $x_1, \ldots, x_m$ and r sinks of demand $y_1, \ldots, y_r$. The cost of transporting material (per unit) from source x_i to sink y_j is d_{ij}. The available supply at a source x_i is a_i and the demand at sink y_j is d_j. It is assumed that the total demand equals the available supplies, i.e.,

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{r} b_j \tag{3.35}$$

We wish to satisfy the demands from the sources in such a way as to minimize the cost of transportation.

For convenience we summarize the initial data in Table 3.1. The

Table 3.1 Unit Transportation Cost d_{ij} from a_i to b_j

	b_1	...	b_r
a_1	d_{11}	...	d_{1r}
$\vdots$	$\vdots$		$\vdots$
a_m	d_{m1}	...	d_{mr}

transport network corresponding to this problem can be constructed as follows. The source s is joined with every vertex x_i by an edge of capacity $c(s, x_i) = a_i$. Every vertex y_i is joined with the sink t by an edge of capacity $c(y_j, t) = b_j$. The cost of sending a flow along edges (s, x_i) and (y_j, t) is assumed equal to zero. Finally, every vertex x_i is connected with every vertex y_j by an edge of infinite capacity and transportation cost (per unit) equal to d_{ij}. This network is amenable to application of the cost optimization method.

Example 3.5. Find the cost-optimal routes for the transportation problem specified by Table 3.2. This table corresponds to the road

Table 3.2 Unit Transportation Cost d_{ij} from a_i to b_j

	5	10	20	15
10	8	3	5	2
15	4	1	6	7
25	1	9	4	3

map shown in Fig. 3.12. The associated transport network is shown in Fig. 3.13*a*.

The partial flows and routes in the order of evaluation on the lines of the above method are as follows

k	1	2	3	4	5	6	
(x_i, y_j) . .	(x_3, y_1)	(x_2, y_2)	(x_1, y_4)	(x_3, y_4)	(x_3, y_3)	(x_2, y_3)	
f_k	5	10	10	5	15	5	50
d_{ji}	1	1	2	3	4	6	—
$d_{ij}f_k$ . . .	5	10	20	15	60	30	140

The total flow is 50 and the total transportation cost is 140. The respective transport flow pattern is shown also in the road map.

It is not hard to see that in general the transportation cost in a network like that is defined by the expression

$$\sum_{i=1}^{m} \sum_{j=1}^{r} d_{ij} f(x_i, y_j) \tag{3.36}$$

This method of solution of the transportation problem is essentially a method of evaluation of partial flows $f(x_i, y_j)$ that minimize the transportation cost (3.36). We shall meet this type of problems later, in Sect. 8.2 devoted to linear programming. There, they will be handled by alternative solution techniques.

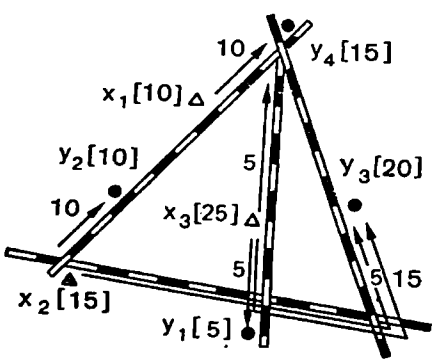


Fig. 3.12. Road map

Let us now turn to a transportation problem which should be optimized subject to a *time criterion*. An example problem will be that given by Table 3.2 where d_{ij} will be treated as time taken to transport goods from x_i to y_j. Problems of this kind arise in transportation of perishable goods or relief supplies to catastrophe areas. Such situations require that all goods should be brought to destinations as fast as possible.

Consider a general approach to the solution of such problems. Assume that some tentative distribution of the flow f_t in the graph G representing the transport network under study is found. Let the graph G' be a partial graph of G that consists solely of the edges that transmit the flow f_t. Let μ be a path from s to t and t_μ be the time taken to send a flow in this path. Clearly the time to transport all goods from s to t will be controlled on the longest path, since transportation

through all the other paths will terminate earlier. Consequently the time taken to transport all goods will be

$$T = \max_{\mu \in G'} t_\mu \tag{3.37}$$

Thus, the solution of the transportation problem reduces to the evaluation of a partial graph G' of G that would be able to pass the flow f and whose longest path would be a minimum of those in all

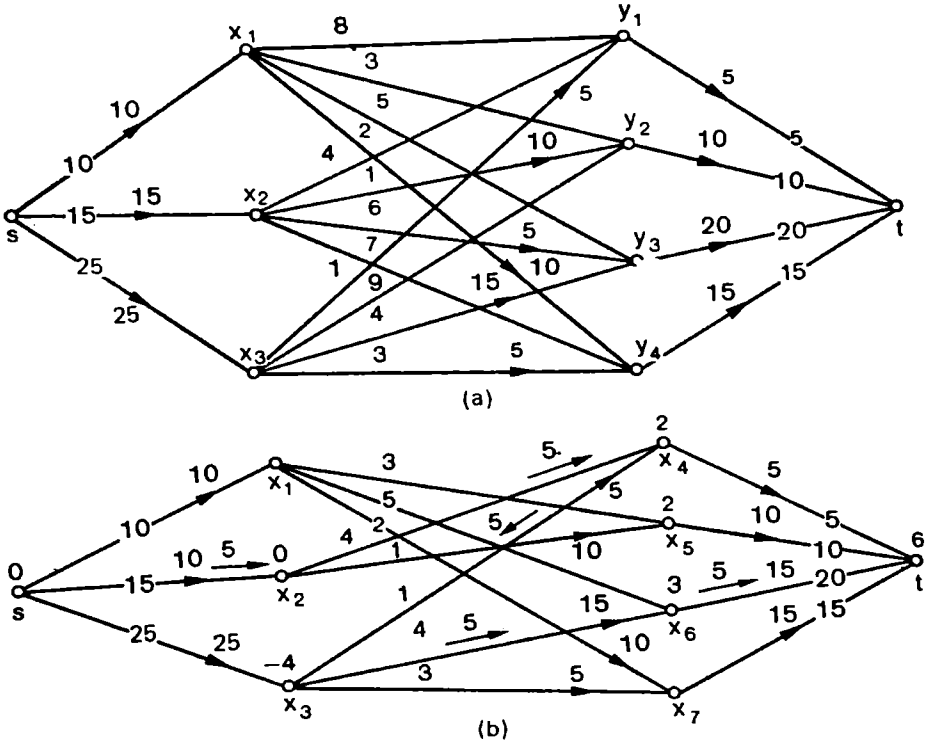


Fig. 3.13. Solution of the transportation problem for (*a*) optimum cost, and (*b*) minimum time

other similar graphs. It is instructive to note that the cost-optimal solution minimizing the expression (3.36) may not be the best one from the optimal-time standpoint.

The solution of this problem boils down to a step-by-step improvement of graph G' by removing from this graph the longest paths and augmenting it by short paths that have not been used.

To illustrate the procedure we turn to the latest example. As a first approximation to the best solution we initialize the graph by the solution obtained for the cost optimization version. The flow pattern for this case is presented in Fig. 3.13*a*. The time to traverse the longest path (x_2, y_3) is seen to be six units. However, shorter routes (x_1, y_3),

(x_2, y_1), and (x_1, y_3) are out of use. Therefore, one of them may help deleating the path (x_2, y_2).

We construct a partial graph G' whose edge set will consist of those edges of G for which $t_\mu < 6$, and in which the flow pattern remains the same as in G. This graph G' is depicted in Fig. 3.13*b*, where for convenience vertices y_i are relabeled as $x_{m+i} = x_{3+i}$. The flow f_t' through this graph is 45 units, i.e. five units less than the initial flow $f_t = 50$ units. This is a full flow, since all the s-t paths in G' involve saturated edges. This flow, however, may not be the maximum one.

If the maximum flow in G' is f_t, then there exists a flow pattern in which the longest path has $t_\mu < 6$. Consequently the further solution reduces to the evaluation of the maximum flow in G'.

Figure 3.13*b* shows a labeling of vertices of G' in agreement with the labeling procedure of Sect. 3.3. This labeling indicates that there is a

Table 3.3

Path	Partial flow	Time
(x_2, y_2)	10	1
(x_1, y_4)	10	2
(x_3, y_4)	5	3
(x_2, y_1)	5	4
(x_3, y_3)	20	4

path $(s, x_2, x_4, x_3, x_6, t)$ in which the s-t flow can be increased by 5 units. This additional flow is indicated by arrows. The maximum flow in G' is $f_t = 50$ units and the longest path takes 4 units. No further reduction of time can be achieved as there are no paths to x_6 whose time is less than 4 units.

The finite distribution of partial flows that represents the solution of this transportation problem for minimum time is summarized in Table 3.3.

4

ELEMENTS OF LINEAR ALGEBRA AND CONVEX SETS

4.1 Metric Spaces and Distances

4.1.1 The Concept of Distance

The concept of sets as collections of isolated elements is of limited value because most of actual agglomerations of objects involve some kind of connections or interactions between them. It would be useful therefore to augment this notion by the relation between elements.

A set in which some relations are specified between its elements or some operations are defined on these elements is called a *space*. To begin with we discuss *metric spaces* which are defined by the distance between elements [14]. As a rule we understand distance as a measure of separation of one thing from another. Geometrically this is the

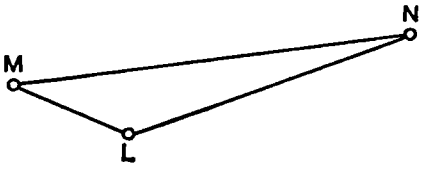


Fig. 4.1. Illustration of the triangle inequlaity

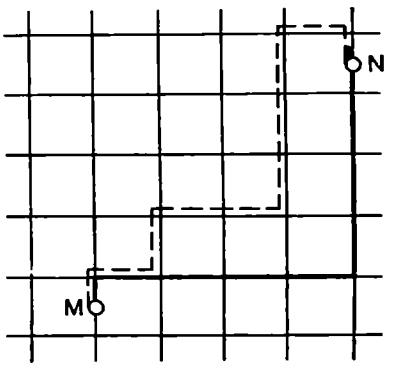


Fig. 4.2. Distance in a town separated into blocks

length of a straight line $d(M, N)$ between points M and N, as illustrated in Fig. 4.1. In everyday life this definition is often insufficient, as it is not always feasible to pass a straight line between two points of interest, as in a town where one can move only in parallel and perpendicular streets (Fig. 4.2). The distance may have some physical background as a waterway distance or railway route.

On the other hand, separation is frequently used without its geometrical meaning of a straight line in space. For example, the separation of two fields on a chessboard can be assessed by the number of moves that should be made to move the figure from one field to the other. Two adjacent fields are in close proximity for a king, but not

for a knight, and may be viewed as infinitely distant for a bishop if they are of different color.

In mathematics one often needs to replace or fit one function $y = f(x)$ with another, more convenient function $y = g(x)$. This substitution is legitimate if $g(x)$ is close to $f(x)$, as illustrated in Fig. 4.3*a*. Of course, we should be definite as to what we mean by saying that the functions are close to one another. If we take the areas confined by these functions for comparison, then for all the situations shown in Fig. 4.3 the functions may be deemed close to each

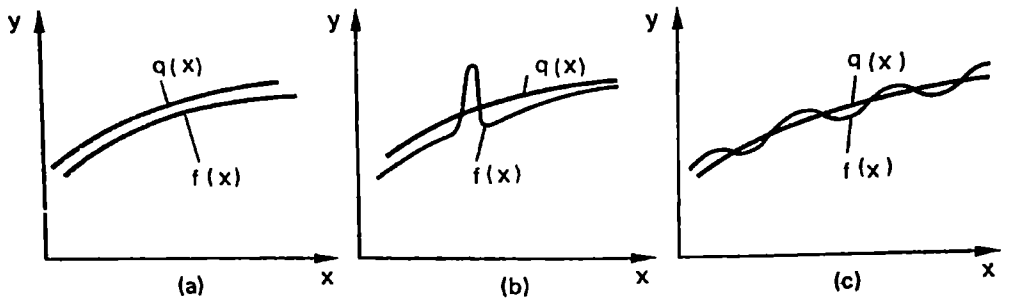


Fig. 4.3. Different situations with two close functions

other. If, on the other hand, we estimate the closeness of the functions $g(x)$ and $f(x)$ by their maximum deviations for every x, then the diagram at (*b*) shows functions that are no longer close. Moreover, if we take into account the behavior of the functions, i.e. require that their derivatives be close as well, then the functions will not meet the closeness criterion in the case of (*c*).

These examples suggest that some general definition should be developed for the distance as a measure of separation of two points, and, consequently, as a measure of space in which these points are embedded. The content of this concept may be different in specific situations. Because the collections of objects are described as sets, the concepts of space and distance should be connected with the concept of a set.

4.1.2 Definition of a Metric Space

The concept of a distance between elements of a set X evolves by generalizing the geometric properties of distance intuitively conceivable from Fig. 4.1.

We associate to each pair of elements in X some real non-negative number $d \geqslant 0$. This number is called a *distance* or *metrix* in X if for any x, y, and z in X it possesses the following three properties:

(1) $d(x, y) = 0$ if and only if $x = y$ (identity axiom);

(2) $d(x, y) = d(y, x)$ (symmetry axiom); and

(3) for any three elements x, y, z in X, $d(x, y) \leqslant d(x, z) + d(z, y)$ (triangle inequality).

A set X with a metric d defined on it is called a *metric space*. Elements of X are referred to as points of the metric space (X, d).

From this definition it follows that a set X becomes a metric space only after a metric $d(x, y)$ has been defined on it. Different metrics defined on a set X result in different spaces. For example, the spaces depicted in Figs. 4.1 and 4.2 consist of points of the same plane, but have different metrics.

4.1.3 Examples of Metric Spaces

Let x and y be elements of the set R of real numbers. The set R becomes a metric space if the distance between x and y is defined by the expression

$$d(x, y) = |x - y| \tag{4.1}$$

This formula is used to determine the distance between points of the real axis that represents a simple example of metric space.

In Sect. 2.3 we have learned that ordered n-element sets are convenient to view as points in an imaginary n-dimensional space R^n. The representation of such a space may be expanded by introducing for it the distance between two points, i.e. by treating this space as a metric one. Denote by $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ two distinct points of R^n. There are several ways to convert R^n into a metric space. More often than not the distance between x and y is defined by the formula

$$d_2(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2} \tag{4.2}$$

For $n = 2$ or 3 this definition coincides with the ordinary geometric concept of distance. The properties (1)-(3) hold true for this distance, as can be clearly seen from Fig. 4.1.

The metric $d_2(x, y)$ is called Euclidean and the space R^n with such a metric is referred to as a Euclidean space and denoted by E^n.

Another way to introduce the concept of distance in R^n is by the formula

$$d_1(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{4.3}$$

or

$$d_\infty(x, y) = \max\left(|x_1 - y_1|, \ldots, |x_n - y_n|\right) \tag{4.4}$$

As can be readily seen the metrics d_2, d_1, and d_∞ are particular cases of the metric

$$d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \tag{4.5}$$

The properties (1) and (2) clearly hold for the metrics (4.3) and (4.4). To prove the property (3) we take another point $z = (z_1, \ldots, \ldots, z_n)$ in R^n. For $d_1(x, y)$ we have

$$d_1(x, y) = \sum_{i=1}^{n} |x_i - z_i + z_i - y_i| \leqslant \sum_{i=1}^{n} (|x_i - z_i| + |z_i - y_i|)$$
$$= d_1(x, z) + d_1(z, y)$$

For the distance $d_\infty(x, y)$ the property (3) is verified by assuming that $|x_k - y_k|$ is the largest of the corresponding differences of x and y. Then

$$d_\infty(x, y) = |x_k - y_k| = |x_k - z_k + z_k - y_k|$$
$$\leqslant |x_k - z_k| + |z_k - y_k|$$

Since

$$|x_k - z_k| \leqslant \max(|x_1 - z_1|, \ldots, |x_n - z_n|) = d_\infty(x, z)$$

and

$$|z_k - y_k| \leqslant \max(|z_1 - y_1|, \ldots, |z_n - y_n|) = d_\infty(z, y)$$

then

$$d_\infty(x, y) \leqslant d_\infty(x, z) + d_\infty(z, y)$$

If we consider the set of all functions of time continuous on the interval $a \leqslant t \leqslant b$ and take $x(t)$ and $y(t)$ to be two functions of the set, then the distance between them is

$$d(x, y) = \max_{a \leqslant t \leqslant b} |x(t) - y(t)| \tag{4.6}$$

It is seen to satisfy all the properties of the metric. We denote the space with this metric by $C_{[a, b]}$.

4.2 Metric Spaces in Cybernetics Problems

4.2.1 Message Space

In the Introduction we learned that messages are transmitted over channels of communication systems with the aid of some alphabet that consists of a finite number of symbols. A message is a string of symbols of a particular alphabet.

Let us consider a situation when an alphabet U_m that consists of m symbols is used to transmit messages n-symbols long. Shorter me-

ssages may be augmented to this length by some symbol, say zero in the binary alphabet. A set of such messages may be viewed as a metric space if a notion of distance between messages is defined.

We shall understand the distance $d(x, y)$ between two messages x and y as the number of positions in which these messages do not coincide. The resulting metric space will be denoted by $E(n, U_m)$ and referred to as a *message space.*

Example 4.1. Let U_m be the English alphabet, $n = 7$, $x = pattern$, and $y = patient$. These words differ in their fourth, sixth, and seventh letters, hence $d(x, y) = 2$.

Example 4.2. Let $U = U_2 = \{0, 1\}$ be the binary alphabet, $n = 10$, $x = 0100111010$, $y = 0010110010$. These words differ in 2nd, 3rd, and 7th places, hence $d(x, y) = 3$.

The above definition of distance in message space possesses all the properties of distance. The identity and symmetry axioms are obvious. Let us check that the triangle inequality holds.

Consider three messages x, y, and z of length n. Denote by x_k, y_k, and z_k the symbols occupying k-th places in these messages. Naturally, if $x_k = y_k$ and $y_k = z_k$, then $x_k = z_k$, that is, if the symbols of x and y and of y and z coincide in some place, then the symbols of x and z occupying this place also coincide. Consequently, x and z will have distinct symbols only where x and y or y and z have distinct symbols. This implies, however, that the total number of distinct symbols of z and x cannot exceed the sum of distinct symbols of x and y and y and z.

By way of simple illustration we take the binary alphabet $U_2 = \{0, 1\}$. This choice does not inflict any loss of generality as messages encoded in symbols of one alphabet can be always recoded into another alphabet. Let, for example, the alternative alphabet U_m consist of m symbols, numbered from 0 to $m - 1$. If we replace in a message represented in U_m the symbols by their numbers expressed in binary representation, the message will be encoded in the binary alphabet.

We shall refer to the set of symbols of the binary alphabet used to denote a message as a *binary code* or *message code.*

4.2.2 Noise-Immune Codes

A message transmitted in a transmission channel can be corrupted by noise. The signals may be changed so that the detector at the receiving end take zero signal for a one-signal and vice versa. It is desirable therefore to work with codes that enable the decoder to readily evaluate distortions suffered by the message in the channel or even restore the message to its original content. This type of codes is said to be called *noise-immune.*

Consider the case when it is known beforehand that distortions cannot corrupt more than k places of a message. In a message space $E(n, U_2)$ we pick up a subset $H_k \subseteq E(n, U_2)$ such that for any x, y in H_k

$$d(x, y) > k \tag{4.7}$$

We refer to this set H_k as a set of *meaningful words*. The term *word* is here synonymic to "message". Then any x not in H_k is a meaningless word. Suppose that a word $x \in H_k$ is corrupted by noise and becomes a word x'. Since not more than k errors can occur in a word, $d(x, x') \leqslant k$ and $x' \notin H_k$, i.e. x' is a *meaningless word*. Thus a reception of a meaningless word implies that a distortion took place in the transmission of this message. This type of codes satisfying (4.7) is called *error-detecting codes*.

Example 4.3. Pick up in a set $E(3, U_2)$ a set of meaningful words satisfying the condition $d(x, y) = 2$, namely,

$$H_1 = \{000, 101, 011, 110\}$$

If an error occurs in any position of these words, they become meaningless and this individual error can be detected.

A set H_1 producing a code amenable to evaluation of individual errors in words of length n may be constructed as follows. Consider a set $E(n-1, U_2)$, i.e. the set of words $n-1$ symbols long. The set H_1 can be obtained by appending this set by one more position whose value may be selected such that the total number of ones in the words $x \in H_1$ be even.

Example 4.4. For $n = 4$,

$$E(3, U_2) = \{000, 001, 010, 100, 011, 101, 110, 111\}$$

and

$$H_1 = \{0000, 0011, 0101, 1001, 0110, 1010, 1100, 1111\}$$

A corrupted word can readily be revealed with the aid of *modulo-2 addition* in which

$$\begin{aligned} 0 \oplus 0 &= 0 \\ 0 \oplus 1 &= 1 \\ 1 \oplus 0 &= 1 \\ 1 \oplus 1 &= 0 \end{aligned} \tag{4.8}$$

By way of example, in a word $x = a_1, a_2, \ldots, a_n$ the quantity $\beta = a_1 \oplus a_2 \oplus \ldots \oplus a_n$ will be a zero or one depending upon whether the number of ones in the word is even or odd.

It will be instructive to note that the four-digit (hexadecimal) error detecting code obtained in Example 4.4 enables $2^4 = 16$ distinct words, though only 8 of them can be transmitted as meaningful words. Codes having the number of meaningful words less than the total num-

ber of words are known as *redundant codes*. Redundancy is a necessary condition for the construction of noise-immune codes.

Codes that enable correction of transmission errors are called *error-correcting codes*. To construct such a code we assume again that the transmission channel cannot introduce more than k errors, i.e. not more than k positions in a word can be corrupted by noise. The set of meaningful words $H_k = E(n, U_2)$ is selected such that for any x, y in H_k

$$d(x, y) > 2k \tag{4.9}$$

Suppose that errors convert a word x into x'. Since $d(x, y) > 2k$ and $d(x, x') \leqslant k$, then from the triangle inequality it follows that

$$d(x', y) > k \tag{4.10}$$

Thus, the distance of the corrupted word x' to x is smaller than to any other meaningful word and we can restore x by choosing the closest meaningful word to x'. All codes satisfying the condition (4.10) are *error-correcting codes*. The practical implementation of error-correcting codes is rather complicated and treated in special literature [15].

4.2.3 Smoothing of Experimental Errors

In many physical experiments some quantity $y(t)$ is measured over some interval of time. As a rule this function is measured at sampling time instants $t_1, \ldots, t_n$. We shall treat a sequence of these discrete values $Y = (y(t_1), \ldots, y(t_n))$ as a point in n-dimensional space.

Actual experimental conditions are usually far from perfect and measurement data often come distorted. Therefore, the measured values of the physical quantity, $x_1, \ldots, x_n$, will differ from the true values, as illustrated in Fig. 4.4. How can an experimenter possessing the measured values x_i, $i = 1, \ldots, n$, establish the true behavior of the function $y(t)$?

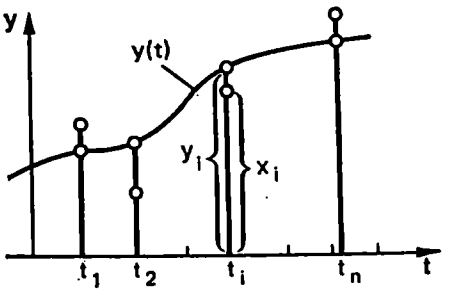


Fig. 4.4. Measurements of a physical quantity

We denote by $X = (x_1, \ldots, x_n)$ a point in n-dimensional space. Clearly, if $X = Y$, i.e. the distance $d(X, Y)$ between X and Y is zero, then the measurements were not corrupted by errors and the sequence $(x_1, \ldots, x_n)$ reflects the true behavior of the physical quantity. If $X \neq Y$, then the degree of distortion may be judged by the distance $d(X, Y)$.

As a rule the general character of $y(t)$ is known accurate to within some numerical parameters whose values would be established by the experimental data. In this case the problem boils down to choosing the unknown parameters of $y(t)$ such that the distance $d(X, Y)$, or what is often more convenient — $d^2(X, Y)$, is a minimum.

We defer the general treatment of such problems until Chapter 10. Here we confine ourselves to discussing the simplest case of constant $y(t) = c$, so that $Y = (c, \ldots, c)$. Making use of the metric (4.2) that defines the Euclidean distance $d_2(X, Y)$ yields for the square of this distance

$$d_2^2(X, Y) = \sum_{i=1}^{n} (x_i - c)^2 \tag{4.11}$$

Equating the derivative of this expression to zero,

$$\frac{\mathrm{d}}{\mathrm{d}c}\left(\sum_{i=1}^{n} (x_i - c)^2\right) = -2\sum_{i=1}^{n} (x_i - c) = 0$$

we have

$$c = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.12}$$

The quantity c obtained with the aid of (4.12) is called the *mean* of the values $x_1, \ldots, x_n$, and the procedure is known as the *method of least squares.*

If the distance is defined differently, then other formulas are used to calculate the mean. For convenience we assume that the quantities $x_1, \ldots, x_n$ are arranged in the order of increasing magnitudes. We leave it to the reader to verify that if the distance is defined by Eq. (4.3), then the mean value, called the *median* in this case, is found as

$$\begin{aligned} c &= x_{(n+1)/2} && \text{for odd } n \\ &= \text{any value from } x_{n(2-1)} \text{ to } x_{n(2+1)} && \text{for even } n \end{aligned} \tag{4.13}$$

To be more specific, at n even it is assumed sometimes that

$$c = 0.5\,(x_{n/2-1} + x_{n/2+1}) \tag{4.14}$$

If the distance is defined by Eq. (4.4), then the formula for the mean value is

$$c = (x_1 + x_n)/2 \tag{4.15}$$

4.2.4 Pattern Recognition

A branch of cybernetics called pattern recognition aims at modeling one of the most valuable properties of human brain — the ability to tell one object, phenomenon, or situation from another, that allows a human being to be at ease in a complicated environment [16, 17].

The major concepts of pattern recognition theory are those of *class* and *image*. Individual objects or phenomena of real life differ from one another by their properties. At the same time they may have some properties in common, that allow us to categorize things or phenomena into sets or classes. For example, various types of cars may be united into a class of automobiles. Classes may be more or less general depending on the properties taken for the categorization. For instance the class of automobiles is an element of a wider class of transportation means. Therefore, a *class* is a set of things or phenomena that have some properties in common.

In any specific situation we usually deal with a finite collection of classes represented by a finite set

$$W = \{A, B, C, \ldots, P\} \tag{4.16}$$

Every class involves a set of objects that may be arbitrarily large in number and diversified in properties. Practical consideration can cover only a limited, and often very small, number of properties. We call the set of these properties of an object a pattern and represent it as an n-vector $x = (x_1, \ldots, x_n)$ whose components are the quantitative characteristics of the properties. An example of a pattern may be a television signal of a photograph divided into n cells. The collection of brightness of these cells forms a point in an n-dimensional space. This point is a pattern of the photograph.

Intuitively one can expect — provided the selected set of features characterizes the properties of the objects appropriately — that the collection of points corresponding to various patterns of the same class will occupy an area in the pattern space, which differs from the areas corresponding to patterns of other classes.

A pattern recognition problem may be deemed solved in principle if the pattern space is divided by surfaces into regions whose points belong to patterns of different classes. One method of solution for this problem consists in a preliminary evaluation of the properties of samples of objects from each class.

Assume that there are only two classes of objects, A and B. Denote by A_0 and B_0 the sets of samples from these classes that have been studied beforehand. The patterns of these samples in the pattern space

may have the form shown in Fig. 4.5*a*, where the points corresponding to samples of A_0 are represented by circles and those of B_0 by triangles.

Let x be the pattern of a new object to be studied. As a measure of closeness of this pattern to classes A and B we take the quantities $S(x, A)$ and $S(x, B)$ representing, say, the mean squares of the di-

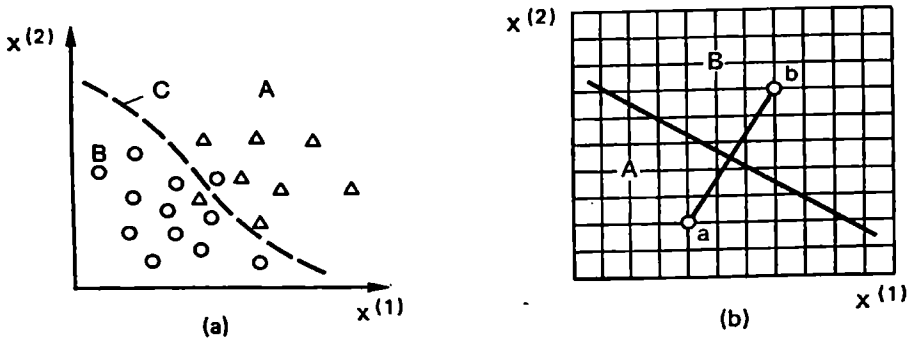


Fig. 4.5. Separation of image space into classes

stance from point x to points corresponding to individual samples of A and B. For example, if $A_0 = \{a_1, \ldots, a_m\}$ and $B_0 = \{b_1, \ldots, b_l\}$, then in accordance with (4.12) we have the mean value in the form

$$\begin{aligned} S(x, A) &= \frac{1}{m} \sum_{i=1}^{m} d^2(x, a_i) \\ S(x, B) &= \frac{1}{l} \sum_{j=1}^{l} d^2(x, b_j) \end{aligned} \tag{4.17}$$

The frontier separating the patterns of both classes in the pattern space will be a surface C whose points satisfy the condition

$$S(x, A) = S(x, B) \tag{4.18}$$

As can be seen from Fig. 4.5*a*, in some cases surface C may happen to be such that an incorrect assignment of a pattern to improper class may occur. A pattern recognition criterion may be a relative number of incorrectly identified patterns.

Example 4.5. With reference to Fig. 4.5*b* we take from the classes A and B samples $a_1 \in A$ and $b_1 \in B$ representing the characteristic features of these classes in the two-dimensional pattern space. A line drawn in this space separates it into two regions corresponding to classes A and B.

We determine the distance by Eq. (4.2) as a line segment joining two points in the pattern space. According to (4.18) the separating line will be the one for which all its points are equidistant from a_1 and b_1. Geometrically, this line is a perpendicular at the middle of the line joining a_1 and b_1.

4.3 Linear Spaces

4.3.1 Definition of a Linear Space

At the outset of this chapter we defined a space as a set with relations specified between its elements. We considered metric spaces specific in assignment to every pair of elements a real number possessing certain properties and called a metric. However, a metric does not exhaust all the structural properties of spaces. Specifically, if a set X consists of real or complex numbers, then the important structural properties include the possibility of obtaining some elements of the set from other by adding these elements or by multiplying an element by a scalar. Sets that possess these properties relate to the class of *linear spaces*. Linear spaces satisfy the following properties.

(1) For each pair of elements x and y in X there is a third element $z \in X$ denoted $x + y$ and called the sum of x and y;

The sum is commutative as $x + y = y + x$ and associative as $x + (y + v) = (x + y) + v$.

There exists in X an element 0 such that $x + 0 = x$ for all x in X.

For every x in X there is an opposite element $-x$ such that $x + (-x) = 0$.

(2) For any number α and every element $x \in X$ there exists an element $\alpha x \in X$, and for any number β

$$(\alpha + \beta)\, x = \alpha x + \beta x \text{ and } \alpha\,(x + y) = \alpha x + \alpha y$$

The conditions (1) and (2) imply that the space possesses *additivity* and *uniformity*.

Sets that allow the operations of addition and multiplication by a scalar are many and diverse. Below we give a few examples of such sets.

A set of real numbers R with addition and multiplication defined in the usual manner.

A set of vectors, i.e. oriented line segments in plane or in three dimensions.

A set of all functions $C_{[a, b]}$ continuous on a given interval $[a, b]$, if for any $x(t)$ and $y(t)$ in $C_{[a, b]}$ $x(t) + y(t)$ is the sum of values of $x(t)$ and $y(t)$ taken at the same values of t, and $\alpha\, x(t)$ is a new function obtained from $x(t)$ by multiplication of all its values by α.

The zero function will be $x(t) = 0$ identically equal to zero on the entire interval $[a, b]$.

A set of all polynomials of degree not exceeding a natural number n.

In what follows we confine our attention to linear spaces whose elements are ordered sequences of n real numbers that will be called *vectors*.

It is instructive to note that any function $x(t)$ may be treated as a vector, if viewed as a set of values at a sequence of instants t, while a polynomial may be converted into a vector whose elements are the coefficients of various powers of this polynomial. Therefore, the term vector is often used to denote elements of a linear space.

4.3.2 Vector Algebra

Up to this point we have assumed that elements of sets can be of any nature. At the remaining portion of this chapter we confine our attention to sets whose elements are vectors, i.e. ordered sequences of n numbers $x^{(1)}, \ldots, x^{(n)}$, which can conditionally be written as a column- or row-vector:

$$\begin{pmatrix} x^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ x^{(n)} \end{pmatrix} = [x^{(i)}]_1^n; \; (x^{(1)}, \ldots, x^{(n)}) = (x^{(i)})_1^n$$

The numbers constituting a vector will be called *elements* or *components* of the vector.

If we denote one representation of a vector, say column vector, by $\mathbf{x}$, then the same vector written in the alternative form, i.e. as a row-vector, is denoted by $\mathbf{x}^T$ and called the *transpose* of $\mathbf{x}$. Thus, the transposition of a vector is the transition from the column-vector to the row-vector form and vice versa. In this chapter we shall be dealing primarily with column-vectors, therefore, if not stated otherwise, vectors will be assumed in the column-vector form written as

$$\mathbf{x} = [x^{(i)}]_1^n = (x^{(1)}, \ldots, x^{(n)})^T \tag{4.19}$$

The number n of components in a vector is called the *dimension* of the vector. Components may be real or complex, but in what follows we confine our consideration to real-valued vectors. A vector of dimension one is simply a real number, i.e. a scalar.

The properties of the discussed ordered sequences of numbers are a generalization of the properties of vectors discussed in elementary algebra.

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are called equal if they have equal components:

$$\mathbf{x} = \mathbf{y} \text{ means } x^{(i)} = y^{(i)}, \; i = 1, \ldots, n \tag{4.20}$$

The *sum* of two vectors $\mathbf{x} + \mathbf{y}$ is a vector whose components are the sums of the respective components in $\mathbf{x}$ and $\mathbf{y}$:

Clearly,

$$\mathbf{x}+\mathbf{y}=\begin{pmatrix} x^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ x^{(n)} \end{pmatrix}+\begin{pmatrix} y^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(n)} \end{pmatrix}=\begin{pmatrix} x^{(1)}+y^{(1)} \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ x^{(n)}+y^{(n)} \end{pmatrix} \tag{4.21}$$

Clearly,

$$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x} \tag{4.22}$$

$$\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z} \tag{4.23}$$

The *difference* of two vectors $\mathbf{x} - \mathbf{y}$ is a vector $\mathbf{z}$ such that $\mathbf{y} + \mathbf{z} = \mathbf{x}$.

Multiplication of a vector $\mathbf{x}$ by a scalar α yields a vector whose components are the components of $\mathbf{x}$ multiplied by α,

$$\alpha\mathbf{x} = \mathbf{x}\alpha = \alpha\begin{pmatrix} x^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ x^{(n)} \end{pmatrix}=\begin{pmatrix} \alpha x^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ \alpha x^{(n)} \end{pmatrix} \tag{4.24}$$

We shall also refer to a vector $\mathbf{x}$ as a *point* of linear space.

4.3.3 Linear Dependence and Independence of Vectors

Let X be a linear space. Consider a set of points in X denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_n$. These points are referred to as *linearly dependent* if there exist numbers $\alpha_1, \ldots, \alpha_n$, not all of which are identically zero, such that

$$\alpha_0\mathbf{x}_1 + \ldots + \alpha_n\mathbf{x}_n = 0 \tag{4.25}$$

In other words, vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly dependent if one of them can be represented as a linear combination of the others. For example, suppose that $\alpha_n \neq 0$, then letting $\alpha_i' = -\alpha_i/\alpha_n$, the condition of linear dependence is

$$\mathbf{x}_n = \alpha_1'\mathbf{x}_1 + \ldots + \alpha_{n-1}'\mathbf{x}_{n-1} \tag{4.26}$$

Example 4.6. In the space constituted by ordered real n-tuples, the linearly independent vectors are

$$\mathbf{x}_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$$
$$\mathbf{x}_2 = (0, 1, \ldots, 0)^{\mathrm{T}}$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$\mathbf{x}_n = (0, 0, \ldots, 1)^{\mathrm{T}}$$

In the space of continuous functions, for a fixed N, functions $f_1(t) = 1, f_2(t) = t, \ldots, f_N(t) = t^{N-1}$ are linearly independent.

4.3.4 Linear Subspace

From the definition of a linear space X it follows that for any points $\mathbf{x}_1$ and $\mathbf{x}_2$ in X and any numbers α and β the quantity $\alpha\mathbf{x}_1 + \beta\mathbf{x}_2$ will be also an element of X.

Consider an arbitrary finite set of points $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ of the linear space X. We wish to demonstrate that the set

$$L(S) = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i \tag{4.27}$$

at any α_i is also a linear set, being a subset of X and accordingly referred to as a *linear subspace.*

Proof. (1) The fact that $L(S)$ is a subset of X follows from the definition of a linear space. Every addend in (4.27) is an element of X. By adding elements of a linear space we again obtain elements of the same linear space X.

(2) To prove that $L(S)$ is a linear space it suffices to demonstrate that for any $\mathbf{x}'$ and $\mathbf{x}'' \in L(S)$ the sum $\alpha\mathbf{x}' + \beta\mathbf{x}''$ also belongs to $L(S)$.

Let us take arbitrary collections of α_i' and α_i'' $(i = 1, \ldots, m)$, then the points

$$\mathbf{x}' = \sum_{i=1}^{m} \alpha_i' \mathbf{x}_i', \quad \mathbf{x}'' = \sum_{i=1}^{m} \alpha_i'' \mathbf{x}_i$$

belong to $L(S)$. Let us consider a point

$$\mathbf{x} = \alpha\mathbf{x}' + \beta\mathbf{x}'' = \alpha \sum_{i=1}^{m} \alpha_i' \mathbf{x}_i + \beta \sum_{i=1}^{m} \alpha_i'' \mathbf{x}_i = \sum_{i=1}^{m} (\alpha\alpha_i' + \beta\alpha_i'')\, \mathbf{x}_i = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i$$

where $\alpha_i = \alpha\alpha_i' + \beta\alpha_i''$. Since this $\mathbf{x}$ is defined by (4.27), it belongs to $L(S)$. Therefore, $L(S)$ is a linear space.

Because the relation (4.27) associates with the set S a linear space $L(S)$, the latter is said to be defined by the set S.

4.3.5 Dimension of a Linear Space

If elements of a set S that define linear space $L(S)$ are linearly dependent, then some of them may be represented as linear combinations of other elements. Deleting these elements of S does not affect the definition of the linear space $L(S)$. Consequently, if we pick up a subset $S_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of S such that all elements in S_n are linearly independent, then the linear space $L(S_n)$ defined by S_n coincides with the linear space $L(S)$. Accordingly, the number of linearly independent elements in set S, i.e. the number n, is called the *dimension of the linear space* $L(S)$, i.e. $L(S)$ is *n-dimensional.*

4.4 Euclidean Spaces

4.4.1 Normed Linear Space

A linear space does not acquire a complete description unless the properties of additivity and uniformity are augmented by the possibility to measure the values of the elements. For example, we cannot compare vectors unless we have agreed in advance what should be taken as the value (length) of a vector. If we introduce the numerical estimates of elements into a linear space, then we obtain the concept of a normed linear space sometimes called a *Banach space.*

A linear space is called a *normed linear space* if it has associated with each element $\mathbf{x}$ in X a real number $\| \mathbf{x} \|$ called the *norm* of $\mathbf{x}$, satisfying the following postulates

$$\| \mathbf{x} \| = 0 \text{ if only if } \mathbf{x} = 0$$

$$\| \alpha \mathbf{x} \| = | \alpha | \cdot \| \mathbf{x} \|$$

$$\| \mathbf{x} + \mathbf{y} \| \leqslant \| \mathbf{x} \| + \| \mathbf{y} \| \text{ (triangle inequality) for all } \mathbf{x} \text{ and } \mathbf{y}$$

It is not hard to establish that the quantity $\| \mathbf{x} - \mathbf{y} \|$ possesses all the properties of distance $d(\mathbf{x}, \mathbf{y})$ in the metric space. Indeed, $\| \mathbf{x} - \mathbf{y} \| = 0$ if $\mathbf{x} - \mathbf{y} = 0$, i.e. if $\mathbf{x} = \mathbf{y}$. Observing that $\mathbf{y} - \mathbf{x} = -(\mathbf{x} - \mathbf{y})$, we find

$$\| \mathbf{y} - \mathbf{x} \| = | -1 | \cdot \| \mathbf{x} - \mathbf{y} \| = \| \mathbf{x} - \mathbf{y} \|$$

$$\| \mathbf{x} - \mathbf{y} \| = \| (\mathbf{x} - \mathbf{z}) + (\mathbf{z} - \mathbf{y}) \| \leqslant \| \mathbf{x} - \mathbf{z} \| + \| \mathbf{z} - \mathbf{y} |$$

Consequently, a normed linear space is a metric space with a metric

$$d(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \| \tag{4.28}$$

All above-mentioned metric spaces when augmented by the properties of additivity and uniformity become normed linear spaces. There exists a special notation for these spaces:

(i) space $C_2^{(n)}$ or E^n is a Euclidean space with the norm

$$\| \mathbf{x} \| = \left(\sum_{i=1}^{n} | x_i |^2 \right)^{1/2} \tag{4.29}$$

which becomes $| \mathbf{x} |$ at $n = 1$;

(ii) space $C_1^{(n)}$ with the norm

$$\| \mathbf{x} \| = \sum_{i=1}^{n} | x_i | \tag{4.30}$$

(iii) space $C^{(n)}$ with the norm

$$\| \mathbf{x} \| = \max \{ | x_1 |, \ldots, | x_n | \} \tag{4.31}$$

(iv) space $C_{[a, b]}$ of continuous functions defined on the interval $[a, b]$ with the norm

$$\| f \| = \max_{a \leqslant t \leqslant b} | f(t) | \tag{4.32}$$

4.4.2 The Scalar Product

We have just learned that a linear space is called Euclidean if the norm of a vector $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^{\mathrm{T}}$ defining its length is calculated by the formula (4.29). This definition of the norm is intimately connected with the concept of scalar (inner or dot) product of vectors. Let

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

be two vectors in a three-dimensional space. The *scalar product* of these vectors is the scalar

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{x} = x_1y_1 + x_2y_2 + x_3y_3 \tag{4.33}$$

The square root of the scalar product of a vector $\mathbf{x}$ by itself is the norm of this vector

$$(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2} = (x_1^2 + x_2^2 + x_3^2)^{1/2} = \| \mathbf{x} \| \tag{4.34}$$

The concepts of vector norm and scalar product of vectors are closely connected to each other, therefore it is important to generalize the concept of this scalar product onto an arbitrary linear space.

Definition. A scalar product is defined on a linear space X if every pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in X is associated with a real number $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ having the following properties:

(1) $\mathbf{x}^{\mathrm{T}}\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{x}$ (commutativity);
(2) $\mathbf{x}^{\mathrm{T}}(\mathbf{y} + \mathbf{z}) = \mathbf{x}^{\mathrm{T}}\mathbf{y} + \mathbf{x}^{\mathrm{T}}\mathbf{z}$ (distributivity);
(3) $\lambda\mathbf{x}^{\mathrm{T}}\mathbf{y} = \lambda(\mathbf{x}^{\mathrm{T}}\mathbf{y})$, for any real λ;
(4) $\mathbf{x}^{\mathrm{T}}\mathbf{x} \geqslant 0$, where $\mathbf{x}^{\mathrm{T}}\mathbf{x} = 0$ only if $\mathbf{x} = 0$.

A linear space in which a scalar product is defined is called a Euclidean space. Some examples of scalar products in various linear spaces follow.

(1) If the vectors are ordered sequences of real numbers with the ordinary definition of addition of vectors and multiplication of a vector by a scalar, then the scalar product of vectors $\mathbf{x} = (x^{(1)}, \ldots, \ldots, x^{(n)})^{\mathrm{T}}$ and $\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})^{\mathrm{T}}$ is defined as

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} x^{(i)}y^{(i)} = \mathbf{y}^{\mathrm{T}}\mathbf{x} \tag{4.35}$$

A mnemonic rule to calculate the scalar product is to multiply the respective elements of the row and column moving along the arrows step by step and to sum up the products, viz.,

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = (\overrightarrow{x^{(1)}, \ldots, x^{(n)}}) \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{pmatrix} \Big| \downarrow$$

(2) If we consider continuous functions $f(t)$ and $g(t)$ given on the interval $[a, b]$ as vectors of space X, then the scalar product of such functions is defined as

$$\int_a^b f(t)\, g(t)\, dt \tag{4.36}$$

It is an easy matter to verify that the relations (4.35) and (4.36) satisfy all the postulates of the scalar product.

The *norm* or *length* of a vector $\mathbf{x}$ in a Euclidean space is the scalar

$$\| \mathbf{x} \| = | \mathbf{x} | = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2} \tag{4.37}$$

This definition satisfies all the conditions of a norm. Indeed, the first two conditions follow from the very definition of a scalar product. Prior to proving the third condition (triangle inequality), we consider the Cauchy inequality, an important inequality in Euclidean spaces, namely,

$$(\mathbf{x}^{\mathrm{T}}\mathbf{y})^2 \leqslant (\mathbf{x}^{\mathrm{T}}\mathbf{x})\ (\mathbf{y}^{\mathrm{T}}\mathbf{y}) \tag{4.38}$$

or

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} \leqslant | \mathbf{x} | \cdot | \mathbf{y} | \tag{4.39}$$

We take the square of the vector $\mathbf{x} + \lambda\mathbf{y}$ to prove the Cauchy inequality, namely, $(\mathbf{x} + \lambda\mathbf{y})^{\mathrm{T}} (\mathbf{x} + \lambda\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{x} + 2\lambda\, (\mathbf{x}^{\mathrm{T}}\mathbf{y}) + \lambda^2\, (\mathbf{y}^{\mathrm{T}}\mathbf{y}) \geqslant 0$. Denoting $a = \mathbf{x}^{\mathrm{T}}\mathbf{x} \geqslant 0$, $b = \mathbf{x}^{\mathrm{T}}\mathbf{y}$, and $c = \mathbf{y}^{\mathrm{T}}\mathbf{y} \geqslant 0$, the above relation can be written as $a + 2\lambda b + \lambda^2 c \geqslant 0$. By completing the expression $a + 2\lambda b = a\,(1 + 2\lambda b/a)$ to a full square, we get

$$a\,(1 + \lambda b/a)^2 + (c - b^2/a)\,\lambda^2 \geqslant 0$$

whence it follows that $b^2 \leqslant ac$, that is, the inequality (4.38) holds true.

Making use of expressions (4.35) and (4.36), we arrive at the following inequalities, being the particular cases of Cauchy's inequality,

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leqslant \left(\sum_{i=1}^{n} x_i^2\right) \left(\sum_{i=1}^{n} y_i^2\right) \tag{4.40}$$

and

$$\left(\int_a^b f(t)\,g(t)\,dt\right)^2 \leqslant \left(\int_a^b f^2(t)\,dt\right)\left(\int_a^b g^2(t)\,dt\right) \tag{4.41}$$

To obtain the triangle inequality we consider the expression

$$(\mathbf{x}+\mathbf{y})^{\mathrm{T}}(\mathbf{x}+\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{x} + 2(\mathbf{x}^{\mathrm{T}}\mathbf{y}) + \mathbf{y}^{\mathrm{T}}\mathbf{y}$$

Since by virtue of Cauchy's inequality $2(\mathbf{x}^{\mathrm{T}}\mathbf{y}) \leqslant 2(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{1/2}$, then

$$\begin{aligned}(\mathbf{x}+\mathbf{y})^{\mathrm{T}}(\mathbf{x}+\mathbf{y}) &\leqslant \mathbf{x}^{\mathrm{T}}\mathbf{x} + 2(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{1/2} + \mathbf{y}^{\mathrm{T}}\mathbf{y} \\ &= [(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2} + (\mathbf{y}^{\mathrm{T}}\mathbf{y})^{1/2}]^2\end{aligned}$$

or

$$|\mathbf{x}+\mathbf{y}| \leqslant |\mathbf{x}| + |\mathbf{y}| \tag{4.42}$$

4.4.3 Angle Between the Vectors. Orthogonal Vectors

Cauchy's inequality has a simple geometric interpretation in a two-dimensional space. Let $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ be two vectors in a plane. Assume that the vector $\mathbf{x}$ coincides with the axis of abscissas so that $x_1 = |\mathbf{x}|$ and $x_2 = 0$, as shown in Fig. 4.6. Denote the angle between $\mathbf{x}$ and $\mathbf{y}$ by α, then

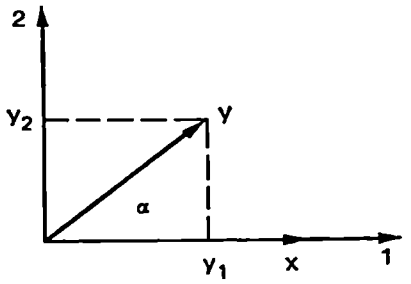


Fig. 4.6. Illustrating a scalar product

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}\mathbf{y} &= x_1y_1 + x_2y_2 \\ &= |\mathbf{x}| \cdot |\mathbf{y}| \cos\alpha\end{aligned} \tag{4.43}$$

whence Cauchy's inequality readily follows by observing that $\cos\alpha \leqslant 1$.

At $|\mathbf{x}| = 1$ this relationship defines the quantity $|\mathbf{y}| \cos\alpha$ being the projection of $\mathbf{y}$ onto the direction of $\mathbf{x}$.

Because the Cauchy inequality has been proved for an arbitrary n-dimensional Euclidean space, then Eq. (4.43) holds in any Euclidean space and the angle between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\alpha = \arccos\frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|} \tag{4.44}$$

At $|\mathbf{x}| = 1$ the scalar product $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ defines the projection of vector $\mathbf{y}$ onto the direction of vector $\mathbf{x}$.

Vectors $\mathbf{x}$ and $\mathbf{y}$ are called *orthogonal* if they are at right angles to each other, i.e. if

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = 0 \tag{4.45}$$

The expression (4.36) extends the concept of scalar product to continuous functions, therefore we shall be justified by saying of orthogonal functions. Functions $f(t)$ and $g(t)$ are orthogonal if

$$\int_a^b f(t)\, g(t)\, \mathrm{d}t = 0 \tag{4.46}$$

4.5 Matrices and Linear Transformations

4.5.1 The Concept of a Matrix

Any rectangular array of numbers (real or complex) with m rows and n columns is termed an $(m \times n)$ *matrix* $\mathbf{A}$

$$\mathbf{A} = \begin{pmatrix} a_{11} \ldots a_{1n} \\ \cdot \quad \cdot \quad \cdot \\ a_{m1} \ldots a_{mn} \end{pmatrix} = [a_{ij}]$$

This is read as m by n matrix. In the special case when m and n are equal, the matrix is said to be *square*. At $m = 1$ the matrix reduces to a row vector, or $1 \times n$ row matrix,

$$(a_1, \ldots, a_n)$$

and at $n = 1$ reduces to a column vector, or $m \times 1$ column matrix,

$$\begin{pmatrix} a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_m \end{pmatrix}$$

Two matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ are equal only when they are of the same size (type) and $a_{ij} = b_{ij}$ for all i and j.

The concept of a matrix arises in connection with linear relations such as linear transformations or systems of linear equations.

4.5.2 Linear Transformations

A transformation of a linear n-dimensional space X is said to be an operator A that maps this space into an m-dimensional linear space Y, or symbolically,

$$A : X \to Y \tag{4.47}$$

In this way a transformation A assigns to each vector $\mathbf{x}$ in X a vector

$$\mathbf{y} = A\mathbf{x} \tag{4.48}$$

in space Y. In a particular case, $Y = X$. Then the transformation assigns to every vector $\mathbf{x}$ in X a vector $A\mathbf{x}$ of the same space.

A transformation A is termed *linear*, if two conditions are satisfied

$$A(\mathbf{x}_1 + \mathbf{x}_2) = A\mathbf{x}_1 + A\mathbf{x}_2$$
$$A(\lambda\mathbf{x}) = \lambda A\mathbf{x} \tag{4.49}$$

These conditions are satisfied if the components $x^{(j)}$ and $y^{(i)}$ of $\mathbf{x}$ and $\mathbf{y}$ are joined by a linear dependence

$$y^{(i)} = \sum_{j=1}^{n} a_{ij}x^{(j)}, \quad i = 1, \ldots, m \tag{4.50}$$

where a_{ij} are arbitrary numbers.

The collection of all numbers a_{ij} for $i = 1, \ldots, m$, and $j = 1, \ldots, n$ forms the matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} \ldots a_{1n} \\ \cdot \cdot \cdot \cdot \cdot \cdot \\ a_{m1} \ldots a_{mn} \end{pmatrix} = [a_{ij}] \tag{4.51}$$

called the *matrix of linear transformation*. From Eq. (4.50) it follows that a linear transformation is completely defined by its matrix A.

Comparing (4.50) and (4.35) reveals that $y^{(i)}$ is the scalar product of row i of A by column-vector $\mathbf{x}$, namely,

$$y^{(i)} = \overrightarrow{(a_{i1}, \ldots, a_{in})} \begin{pmatrix} x^{(1)} \\ \cdot \\ \cdot \\ x^{(n)} \end{pmatrix} \Bigg\downarrow, \quad i = 1, \ldots, m \tag{4.52}$$

The whole set of relations (4.52) can be written as the vector-matrix relation $\mathbf{y} = \mathbf{Ax}$ which leads us to define the matrix-vector product in the form

$$\begin{pmatrix} y^{(1)} \\ \cdot \\ \cdot \\ y^{(m)} \end{pmatrix} = \overrightarrow{\begin{pmatrix} a_{11} \ldots a_{1n} \\ \cdot \cdot \cdot \cdot \cdot \cdot \\ a_{m1} \ldots a_{mn} \end{pmatrix}} \begin{pmatrix} x^{(1)} \\ \cdot \\ \cdot \\ x^{(n)} \end{pmatrix} \Bigg\downarrow \tag{4.53}$$

Comparison of (4.53) and (4.52) indicates that the product of matrix A by column-matrix $\mathbf{x}$ exists only if the number of columns in A is equal to the number of elements in $\mathbf{x}$. The ith element in the resultant column-matrix $\mathbf{y}$ is the scalar product of the i-th row of A and column-matrix $\mathbf{x}$.

Two linear transformations of a linear space X are of fundamental importance,

a zero transformation that brings in correspondence with a vector $\mathbf{x}$ a zero vector $\mathbf{0}$

$$0\mathbf{x} = \mathbf{o} \tag{4.54}$$

a unit transformation that puts in correspondence with each vector x this same vector

$$\mathbf{Ix} = \mathbf{x} \tag{4.55}$$

4.5.3 Matrix Algebra

(i) *Scalar multiplication.* Let **A** be a matrix of a linear transformation **Ax**, and α be a scalar, i.e. a number (when we are simultaneously concerned with numbers and other mathematical objects, such as matrices and vectors, it is customary to refer to numbers as scalars). If $\mathbf{A} = [a_{ij}]$ is an $m \times n$ matrix, then the product $\alpha\mathbf{A}$ is defined as a matrix whose coefficients are αa_{ij}, that is,

$$\alpha\mathbf{A} = [\alpha a_{ij}] \tag{4.56}$$

This implies that the linear transformation α (**Ax**) will be described by the matrix $[\alpha a_{ij}]$.

(ii) *Addition.* Let $\mathbf{y} = \mathbf{Ax}$ and $\mathbf{v} = \mathbf{Bx}$ be two linear transformations with $m \times n$ matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$. Consider a new linear transformation that puts in correspondence with each vector **x** in X a vector $\mathbf{y} + \mathbf{v}$ in Y, namely,

$$\mathbf{y} + \mathbf{v} = \mathbf{Ax} + \mathbf{Bx} = (\mathbf{A} + \mathbf{B})\,\mathbf{x} \tag{4.57}$$

The transformation $(\mathbf{A} + \mathbf{B})\mathbf{x}$ is called the sum of linear transformations **Ax** and **Bx**. To find the matrix $\mathbf{A} + \mathbf{B}$ we represent this transformation in the form of (4.50)

$$y^{(i)} + v^{(i)} = \sum_{j=1}^{n} a_{ij}x^{(j)} + \sum_{j=1}^{n} b_{ij}x^{(j)} = \sum_{j=1}^{n} (a_{ij} + b_{ij})\,x^{(j)}, \quad i = 1, \ldots, m$$

or

$$\mathbf{A} + \mathbf{B} = [a_{ij}] + [b_{ij}] = [a_{ij} + b_{ij}] \tag{4.58}$$

Thus the matrices can be added or subtracted if and only if they are of the same type, i.e. both are $m \times n$ matrices. The elements of the resulting matrix are the sums of the respective elements of the added matrices. From the above definition it follows that

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \tag{4.59}$$

(ii) *Matrix multiplication.* Let X, Y, and Z be linear spaces of dimension m, r, and n, respectively, and let $\mathbf{y} = \mathbf{Bx}$ and $\mathbf{z} = \mathbf{Ay}$ be linear transformations of space X into Y and space Y into Z, where $\mathbf{B} = [b_{kj}]$ and $\mathbf{A} = [a_{ik}]$ are $m \times r$ and $r \times n$ matrices. The product of the transformations **Ay** and **Bx** is called a new linear transformation **Cz** that is achieved by successful transformations $\mathbf{y} = \mathbf{Bx}$ and then $\mathbf{z} = \mathbf{Ay}$. In other words, for any **x** in X we have

$$\mathbf{z} = \mathbf{Cx} = \mathbf{A}\,(\mathbf{Bx}) = \mathbf{ABx} \tag{4.60}$$

The $m \times n$ matrix $\mathbf{C} = \mathbf{AB}$ is called the product of $\mathbf{A}$ by $\mathbf{B}$.

To deduce the procedure of matrix multiplication we represent (4.60) in a linear form. The linear transformations $\mathbf{y} = \mathbf{Bx}$ and $\mathbf{z} = \mathbf{Ay}$ may be written as

$$y^{(k)} = \sum_{j=1}^{m} b_{kj} x^{(j)}, \quad k = 1, \ldots, r$$

$$z^{(i)} = \sum_{k=1}^{r} a_{ik} y^{(k)}, \quad i = 1, \ldots, n$$

Eliminating $y^{(k)}$ yields

$$z^{(i)} = \sum_{k=1}^{r} a_{jk} \sum_{j=1}^{m} b_{kj} x^{(j)} = \sum_{j=1}^{m} \left(\sum_{k=1}^{r} a_{ik} b_{kj} \right) x^{(j)}, \quad i = 1, \ldots, n \tag{4.61}$$

On the other hand, the transformation $\mathbf{z} = \mathbf{Cx}$ can be presented as

$$z^{(i)} = \sum_{j=1}^{m} c_{ij} x^{(j)}, \quad i = 1, \ldots, n \tag{4.62}$$

where c_{ij} are the elements of matrix $\mathbf{C}$.

Comparing (4.61) and (4.62) leads us to the rule that can be used in calculating elements of $\mathbf{C} = \mathbf{AB}$, namely,

$$c_{ij} = \sum_{k=1}^{r} a_{ik} b_{kj}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m \tag{4.63}$$

Thus, the element c_{ij} of matrix $\mathbf{C}$ is the scalar product of the i-th row of $\mathbf{A}$ and j-th column of $\mathbf{B}$, so that symbolically the matrix product $\mathbf{AB}$ may be represented in the form

$$\mathbf{AB} = \begin{pmatrix} \overrightarrow{a_{11} \ldots a_{1r}} \\ \cdots\cdots \\ a_{n1} \ldots a_{nr} \end{pmatrix} \begin{pmatrix} b_{11} \ldots b_{1m} \\ \cdots\cdots \\ b_{r1} \ldots b_{rm} \end{pmatrix} \Bigg| \downarrow \tag{4.64}$$

It is important to observe that the matrix product can be defined only when the number of columns in the first factor is equal to the number of rows in the second factor. Therefore, the existence of the product $\mathbf{AB}$ by no means implies that $\mathbf{BA}$ exists. Moreover, even if $\mathbf{BA}$ does exist, then in general

$$\mathbf{AB} \neq \mathbf{BA} \tag{4.65}$$

Equality in this case takes place only for square matrices, and when $\mathbf{AB} = \mathbf{BA}$ the square matrices are said to be *commuting*.

The following laws of elementary algebra remain valid for the product of matrices:

$$(\mathbf{A} + \mathbf{B})\,\mathbf{C} = \mathbf{AC} + \mathbf{BC} \tag{4.66}$$

$$\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB} \tag{4.67}$$

$$(\mathbf{AB})\,\mathbf{C} = \mathbf{A}(\mathbf{BC}) \tag{4.68}$$

4.5.4 The Transpose of a Matrix

Let $\mathbf{A} = [a_{ij}]$ be an $m \times n$ matrix. The matrix $\mathbf{A}^{\mathrm{T}} = [a'_{ij}]$ on n by m type whose rows are the columns of $\mathbf{A}$ and whose columns are the rows of $\mathbf{A}$ is called the *transpose* of matrix $\mathbf{A}$. The elements a'_{ij} of $\mathbf{A}^{\mathrm{T}}$ can be defined by the elements a_{ij} of $\mathbf{A}$ as

$$a'_{ij} = a_{ji} \tag{4.69}$$

If $\mathbf{A}$ and $\mathbf{B}$ are two $m \times n$ matrices, then it follows from the definition of the transpose of a matrix that

$$\begin{aligned} (\mathbf{A} + \mathbf{B})^{\mathrm{T}} &= \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}} \\ (\mathbf{AB})^{\mathrm{T}} &= \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} \\ (\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} &= \mathbf{A} \end{aligned} \tag{4.70}$$

The concept of the transpose of a matrix leads us to an important relation of matrix theory. Let $\mathbf{A}$ be an $m \times n$ matrix which specifies in an n-dimensional linear space a linear transformation $\mathbf{y} = \mathbf{Ax}$. Consider in this space a vector $\mathbf{z}$ that can be transformed into a vector $\mathbf{v} = \mathbf{A}^{\mathrm{T}}\mathbf{z}$ by means of matrix $\mathbf{A}^{\mathrm{T}}$. Determine the scalar products

$$\mathbf{y}^{\mathrm{T}}\mathbf{z} = (\mathbf{Ax})^{\mathrm{T}}\mathbf{z} \text{ and } \mathbf{x}^{\mathrm{T}}\mathbf{v} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{z} \tag{4.71}$$

In view of (4.70) we have

$$(\mathbf{Ax})^{\mathrm{T}}\mathbf{z} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{z} \tag{4.72}$$

In other words, from the point of view of the scalar product the transformation of vector $\mathbf{x}$ by matrix $\mathbf{A}$ is equivalent to the transformation of vector $\mathbf{z}$ by means of matrix $\mathbf{A}^{\mathrm{T}}$.

4.6 Square Matrices

4.6.1 Zero and Unit Matrices

In the preceding sections of this chapter we have considered the linear transformations mapping a linear space X onto another linear space Y, which are described by rectangular matrices. Here we shall learn additional properties of linear transformations defined on one

and the same space X, i.e. those transforming X onto X. Such transformations are described by square matrices which are of fundamental significance for many applications.

If an $m \times r$ matrix is multiplied by an $r \times n$ matrix, the resulting $m \times n$ matrix will be of the type different from either of the factors. Two $m \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$, called square matrices of order m, can be multiplied in any order, obtaining products $\mathbf{AB}$ or $\mathbf{BA}$. These products are in general not equal but have the same order m that the factors.

A square matrix may be multiplied by itself $\mathbf{AA} = \mathbf{A}^2$. By continuing this process we can define $\mathbf{A}^n$ for integral positive n. Letting $\mathbf{B} = \mathbf{A}^n$, we can define $\mathbf{B}^{1/n} = \mathbf{A}$, that is, obtain $\mathbf{A}^r$ for any positive rational exponent r.

If we do not require that $\mathbf{AB} = \mathbf{BA}$, then rules of elementary algebra of scalars apply to square matrices as well, for example,

$$(\mathbf{A} + \mathbf{B})^2 = \mathbf{A}^2 + (\mathbf{AB} + \mathbf{BA}) + \mathbf{B}^2$$
$$(\mathbf{A} + \mathbf{B})(\mathbf{A} - \mathbf{B}) = \mathbf{A}^2 - \mathbf{B}^2 + (\mathbf{BA} - \mathbf{AB})$$

The matrices associated with the zero transformation (4.54) and unit transformation (4.55) of a linear space X play a key role in the family of square matrices.

A matrix whose elements are all zero is called a *zero* or *null matrix* and is denoted by 0.

If a square matrix $\mathbf{A} = [a_{ij}]$ is such that $a_{ij} = 0$ for all $i \neq j$, that is all the elements both above and below the principal diagonal are zero, then the matrix is called a *diagonal matrix.*

A diagonal matrix whose non-zero elements all equal unity is called *unit matrix* and denoted by I.

The following properties of the unit and zero matrices are self-evident:

$$\begin{aligned} \mathbf{A} + 0 &= \mathbf{A} \\ 0\mathbf{A} &= \mathbf{A}0 = 0 \\ \mathbf{IA} &= \mathbf{AI} = \mathbf{A} \end{aligned} \tag{4.73}$$

Many relations of elementary algebra hold for square matrices, as

$$(\mathbf{I} + \mathbf{A})^n = \mathbf{I} + n\mathbf{A} + \frac{n(n-1)}{2!}\mathbf{A}^2 + \ldots + \mathbf{A}^n \tag{4.74}$$

Nonalgebraic functions of A also occur in applications, for example, the matrix exponent

$$e^{t\mathbf{A}} = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2!}\mathbf{A}^2 + \ldots \tag{4.75}$$

plays an important role in the theory of differential equations.

4.6.2 Determinant of a Square Matrix

The determinant of a square matrix $\mathbf{A} = [a_{ij}]$ will be formally introduced for the third order determinant, $\det \mathbf{A}$, namely,

$$\det \mathbf{A} = \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

The determinant of a matrix $\mathbf{A}$ has the following properties [19]:

(1) if one column of $\mathbf{A}$ is multiplied by a fixed scalar λ, then the value of the determinant $\det \mathbf{A}$ is multiplied by λ;

(2) if two columns are interchanged, then the value of the determinant is multiplied by -1;

(3) if any two columns of $\mathbf{A}$ are equal, then $\det \mathbf{A} = 0$;

(4) if we add to one column of $\mathbf{A}$ any other column multiplied by a scalar factor, then the value of the determinant of $\mathbf{A}$ remains the same;

(5) if columns of $\mathbf{A}$ are linearly dependent, then $\det \mathbf{A} = 0$.

The properties (1)-(4) relate to determinants of any order. Therefore we pause here to prove only item (5).

Proof. We represent the matrix $\mathbf{A}$ as a vector $\mathbf{A} = (\mathbf{A}^1, \ldots, \mathbf{A}^m)$, where $\mathbf{A}^j = (a_{1j}, \ldots, a_{mj})$ for $j = 1, \ldots, m$ denote the columns of $\mathbf{A}$. Accordingly, the determinant of $\mathbf{A}$ will be written as $\det \mathbf{A} = D(\mathbf{A}^1, \ldots, \mathbf{A}^m)$.

If the columns $\mathbf{A}^j$ of $\mathbf{A}$ are linearly dependent, then there exist numbers α_j, $j = 1, \ldots, m$, not all simultaneously zero, such that

$$\sum_j \alpha_j \mathbf{A}^j = 0 \tag{4.76}$$

Let $\alpha_k \neq 0$, then resolving Eq. (4.76) for $\mathbf{A}^k$ and denoting $\mu_j = -\alpha_j/\alpha_k$, we have

$$\mathbf{A}^k = \sum_{j \neq k} \mu_j \mathbf{A}^j$$

Now the determinant is

$$\det \mathbf{A} = D(\ldots \sum_{j \neq k} \mu_j \mathbf{A}^j \ldots) = \sum_{j \neq k} \mu_j D(\ldots \mathbf{A}^j \ldots)$$

In the determinants under the summation symbol, columns $\mathbf{A}^j$, $j = 1, \ldots, m$, $j \neq k$, appear in the place of the column $\mathbf{A}^k$. In other words, each of them has two identical columns and by rule (3) is equal to zero. Hence, $\det \mathbf{A} = 0$ if columns of $\mathbf{A}$ are linearly dependent.

4.6.3 The Inverse Matrix and Solution of Systems of Linear Equations

Consider a linear transformation $\mathbf{y} = \mathbf{Ax}$ with a square matrix $\mathbf{A} = [a_{ij}]$, which transforms a vector $\mathbf{x}$ of space X into a vector $\mathbf{y}$ that is also in X. The *inverse transformation* denoted by $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ is the one that transforms $\mathbf{y}$ into the original vector $\mathbf{x}$. The matrix $\mathbf{A}^{-1}$ of this transformation is said to be *inverse* with respect to the matrix $\mathbf{A}$.

The concept of an inverse matrix can be approached from the standpoint of solution of systems of linear equations. Let us assume that we have a system of n linear equations in n unknowns $(x_1, \ldots, x_n)$

$$\sum_{j=1}^{n} a_{ij}x_j = y_i, \quad i = 1, \ldots, n \tag{4.77}$$

which can be written in matrix form as

$$\mathbf{Ax} = \mathbf{y} \tag{4.78}$$

By definition of the inverse matrix, the solution of this sytem is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} \tag{4.79}$$

This solution can be found by Cramer's rule [19] in terms of determinants as

$$x_j = \frac{1}{\Delta} \sum_{i=1}^{n} C_{ij}y_i, \quad j = 1, \ldots, n \tag{4.80}$$

where $\Delta = \det \mathbf{A}$ is the determinant of matrix $\mathbf{A}$, and C_{ij} is the cofactor of the element a_{ij} in the expansion of the n-th order determinant $\det \mathbf{A}$ by the j-th column

$$\det \mathbf{A} = \sum_{i=1}^{n} a_{ij}C_{ij}$$

Thus the elements of the matrix $\mathbf{A}^{-1}$ are C_{ij}/Δ.

Alternatively, this solution of the given system of n linear equations in n unknowns may be written as $x_j = \Delta_j/\Delta$, $j = 1, \ldots, n$, where Δ_j is the determinant obtained from Δ by replacing the elements in the j-th column of Δ by the elements $y_1, \ldots, y_n$, and where it has been assumed that $\Delta \neq 0$.

An n by n matrix $\mathbf{A}$ is said to be *regular* if its n columns (or rows) are linearly independent when its $\det \mathbf{A} \neq 0$. Such a matrix is called *non-singular*, otherwise it is singular. Likewise the system of equations (4.78) is called *regular* (or independent or non-singular) if its coefficient matrix is regular, otherwise the system is called singular

or dependent. If A is a regular matrix, then the system $\mathbf{Ax} = \mathbf{y}$ has the unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$.

Thus the inverse matrix $\mathbf{A}^{-1}$ exists only if A is non-singular. Suppose that $\mathbf{A} = [a_{ij}]$ is a square matrix of order n, then the transpose of the matrix of the cofactors of the elements of A is called the adjoint of A and is denoted by adj A. In other words, if

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & & a_{nn} \end{pmatrix}$$

and $\mathbf{A}_{ij}$ denotes the cofactor of the element a_{ij}, then the adjoint of the matrix A is the matrix obtained by replacing each element a_{rs} (in row r and column s) by the cofactor of the element a_{sr} (in row s and column r), i.e. by $\mathbf{A}_{sr}$. Now

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \operatorname{adj} \mathbf{A} \tag{4.81}$$

It should be noted that the expression (4.79) may be obtained in a formal way merely by premultiplying both sides of (4.78) by $\mathbf{A}^{-1}$, namely,

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{y} \tag{4.82}$$

Comparing (4.82) and (4.79) yields

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{4.83}$$

The formula (4.81) has the merit of furnishing an explicit expression for the inverse. It is indeed used to calculate inverse matrices of small dimension. However, as the size of a matrix increases this procedure becomes extremely laborious, therefore more efficient procedures are preferred such as Gaussian algorithms [18]. In addition, computer libraries usually provide a subroutine for matrix inversion.

Example 4.7. Given $\mathbf{A} = \begin{pmatrix} 8 & 6 \\ 4 & 5 \end{pmatrix}$, then $\det \mathbf{A} = 8 \times 5 - 4 \times 6 = 16$. Here

$$\mathbf{A}^{-1} = \frac{1}{16}\begin{pmatrix} 5 & -4 \\ -6 & 8 \end{pmatrix}^{\mathrm{T}} = \frac{1}{16}\begin{pmatrix} 5 & -6 \\ -4 & 8 \end{pmatrix}$$

$$= \begin{pmatrix} 0.3125 & -0.375 \\ -0.25 & 0.5 \end{pmatrix}$$

If **A** and **B** are square matrices, then

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

To prove, we premultiply **AB** by $\mathbf{B}^{-1}\mathbf{A}^{-1}$, then

$$\mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{AB} = \mathbf{B}^{-1}(\mathbf{A}^{-1}\mathbf{A})\,\mathbf{B} = \mathbf{I}$$

The concept of the rank of a matrix is closely connected with the inverse of a matrix. The rank is usually defined by means of *minors* of elements, which are the determinants of the next lower order obtained by striking out the row and column of the determinant in which the element lies.

The *rank* of a matrix has the order of the nonzero minor of greatest order that can be selected from the matrix. If r is the rank of an $m \times n$ matrix **A**, then $r \leqslant \min(m, n)$.

A square matrix is non-singular only when its rank is equal to its order, and consequently in this case it has an inverse.

4.6.4 Invariant Subspace. Eigenvectors and Eigenvalues

Let X be an n-dimensional linear space, and $\mathbf{y} = \mathbf{Ax}$ be a linear transformation on X. Let $X_1 \subset X$ be a subspace in X such that if $x \in X_1$, then $\mathbf{y} = \mathbf{Ax} \in X_1$. A subspace with this property is said to be *invariant* with respect to the linear transformation $\mathbf{y} = \mathbf{Ax}$.

One-dimensional invariant spaces, which are straight lines through the origin of X, are of especial interest. If **x** is a point in X and α is a real variable varying from $-\infty$ to $+\infty$, then $\alpha\mathbf{x}$ is a one-dimensional subspace of X that passes through **x** (at $\alpha = 1$) and through the origin (at $\alpha = 0$), as shown in Fig. 4.7 for $n = 2$. This one-dimensional subspace will be denoted R_1.

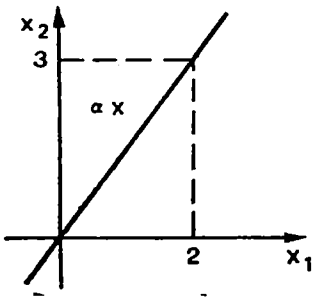


Fig. 4.7. One-dimensional invariant subspace

Suppose that in the infinite set of one-dimensional spaces R_1 there exist spaces R_1 that are invariant with respect to the transformation $\mathbf{y} = \mathbf{Ax}$, i.e. for any **x** in R_1 their is $\mathbf{y} = \mathbf{Ax}$ that also belongs to R_1. Denote by λ the ratio y/x, which can be a real number, i.e. $\mathbf{y} = \lambda\mathbf{x}$. Thus, if R_1 in an invariant subspace, then for **x** in R_1 we have

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{4.84}$$

A vector $\mathbf{x} \neq 0$ satisfying (4.84) is called the *eigenvector* of the matrix **A**, and λ is called the *eigenvalue* of **A**.

To determine the eigenvalues of a matrix we rewrite (4.84) observing that $\mathbf{x} = \mathbf{Ix}$, then

$$(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = 0 \tag{4.85}$$

This relation is a system of linear homogeneous equations that can be rewritten as

$$\begin{aligned}
(a_{11} - \lambda)\, x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0 \\
a_{21}x_1 + (a_{22} - \lambda)\, x_2 + \ldots + a_{2n}x_n &= 0 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\
a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn} - \lambda)\, x_n &= 0
\end{aligned} \tag{4.86}$$

Suppose that we have assumed a numerical value for λ and wish to know whether or not the system (4.86) has a non-trivial (non-zero) solution, i.e. there exist nonzero $x_1, x_2, \ldots, x_n$ that make (4.86) to be identically zero.

To begin with we consider the case of two variables

$$\begin{aligned}
(a_{11} - \lambda)\, x_1 + a_{12}x_2 &= 0 \\
a_{21}x_1 + (a_{22} - \lambda)\, x_2 &= 0
\end{aligned} \tag{4.87}$$

Assuming that both x_1 and $x_2 \neq 0$ we find from the first equation

$$x_2 = -(a_{11} - \lambda)\, x_1/a_{12}$$

The second equation yields after elimination of x_2

$$[a_{21} - (a_{22} - \lambda)\,(a_{11} - \lambda)/a_{12}]\, x_1 = 0$$

This expression has sense provided that

$$(a_{11} - \lambda)\,(a_{22} - \lambda) - a_{12}a_{22} = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0$$

that is when $\det (\mathbf{A} - \lambda\mathbf{I}) = 0$. This is a necessary and sufficient condition for solvability of the system (4.87). If this condition is satisfied, then Eqs. (4.87) yield only the ratio

$$x_1/x_2 = -a_{12}/(a_{11} - \lambda) = -(a_{22} - \lambda)/a_{21} \tag{4.88}$$

that is, the variables themselves can be determined only accurate to a constant factor.

A similar situation takes place with the system (4.85) which has a non-trivial solution only if

$$\det (\mathbf{A} - \lambda\mathbf{I}) = 0 \tag{4.89}$$

The variables, i.e. the vector **x**, are again determined accurate to a constant multiplier.

Equation (4.89) is called the *characteristic equation* of matrix **A**. If the determinant of this equation is expanded, a polynomial in λ, of degree n, is obtained. This is the *characteristic polynomial* corresponding to the matrix **A**, namely,

$$\det (\mathbf{A} - \lambda\mathbf{I}) = q(\lambda) = q_0 + q_1\lambda + \ldots + q_n\lambda^n$$

It is an easy matter to see that $q_0 = \det \mathbf{A}$ and $q_n = (-1)^n$. Hence the characteristic equation may be written

$$q_0 + q_1\lambda + \ldots + q_n\lambda^n = 0 \tag{4.90}$$

This is a polynomial equation of degree n, and so it has n roots, not all necessarily distinct, which are called the *eigenvalues* of the matrix **A**. These eigenvalues must not be all real, but since **A** is a real-valued matrix, the complex roots will occur in conjugate pairs.

Substituting any eigenvalue λ_i into the initial system (4.85), we obtain the equation

$$(\mathbf{A} - \lambda_i\mathbf{I})\,\mathbf{x} = 0$$

that has a non-trivial solution since $\det(\mathbf{A} - \lambda_i\mathbf{I}) = 0$. This solution is given by a vector $\mathbf{x}^i$ defined accurate to a scalar multiplier. This solution corresponds to λ_i and is called the *eigenvector* of the matrix **A** corresponding to λ_i.

Since λ_i may be a complex number, $\mathbf{x}^i$ may contain complex components. However, complex roots occur in pairs, therefore complex-valued eigenvectors will also occur in pairs.

In accordance with (4.84) eigenvectors $\mathbf{x}^i$ corresponding to eigenvalues λ_i are related as

$$\mathbf{A}\mathbf{x}^i = \lambda_i\mathbf{x}^i,\ i = 1, \ldots, n \tag{4.91}$$

or in a compact form

$$\mathbf{AV} = \mathbf{V}\Lambda \tag{4.92}$$

where **V** is a square matrix constituted by the eigenvectors of the matrix **A**, and Λ is a diagonal matrix whose main diagonal is occupied by λ_i, all the other entries being zero,

$$\mathbf{V} = (\mathbf{x}^1, \ldots, \mathbf{x}^n) = \begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_n^1 & \ldots & x_n^n \end{pmatrix} \tag{4.93}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & & \lambda_n \end{pmatrix} \tag{4.94}$$

4.6.5 Diagonalization of Matrices

The form of a square matrix **A** of a linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$ can be changed without changing the characteristic equation of this matrix, by means of the *similarity transformation*

$$\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C} \tag{4.95}$$

where $\mathbf{C}$ is a non-singular matrix.

This transformation yields a matrix $\mathbf{B}$ that satisfies the same characteristic equation as the matrix $\mathbf{A}$ and consequently has the same eigenvalues. Indeed, it is an easy matter to see that

$$\mathbf{C}^{-1}(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{C} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C} - \lambda\mathbf{C}^{-1}\mathbf{I}\mathbf{C} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C} - \lambda\mathbf{C}^{-1}\mathbf{C}\mathbf{I} = \mathbf{B} - \lambda\mathbf{I}$$

It is easy to prove that $\det \mathbf{AB} = \det \mathbf{A} \det \mathbf{B}$, then

$$\det(\mathbf{B} - \lambda\mathbf{I}) = \det[\mathbf{C}^{-1}(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{C}] = \det \mathbf{C}^{-1} \det(\mathbf{A} - \lambda\mathbf{I}) \det \mathbf{C}$$

Since the matrix $\mathbf{C}$ is non-singular, from $\det(\mathbf{B} - \lambda\mathbf{I}) = 0$ it follows that $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$.

The similarity transformation reduces some types of square matrices to a more convenient *diagonal form*. The following theorem on the linear independence of eigenvectors of a matrix helps to decide whether or not a matrix can be reduced to a diagonal form.

Theorem 4.1. If the first k eigenvalues of the matrix $\mathbf{A}$ of order n, $\lambda_1, \ldots, \lambda_k$, are distinct, then the corresponding eigenvectors $\mathbf{x}^1, \ldots, \mathbf{x}^k$ are linearly independent.

Proof. We proceed by induction. For $k = 1$ the statement is obvious. Assume that this statement is true for $k - 1$ vectors, i.e. that the eigenvectors $\mathbf{x}^1, \ldots, \mathbf{x}^{k-1}$ are linearly independent, and demonstrate that it is true for all k vectors.

Suppose the contrary, i.e. that the vectors $\mathbf{x}^1, \ldots, \mathbf{x}^k$ are linearly dependent, and

$$\alpha_1\mathbf{x}^1 + \alpha_2\mathbf{x}^2 + \ldots + \alpha_k\mathbf{x}^k = 0$$

where not all α_i are simultaneously zero. We suppose that $\alpha_1 \neq 0$ and subject this sum to the linear transformation defined by the matrix $\mathbf{A}$,

$$\mathbf{A}(\alpha_1\mathbf{x}^1 + \alpha_2\mathbf{x}^2 + \ldots + \alpha_k\mathbf{x}^k) = 0$$

or

$$\alpha_1\lambda_1\mathbf{x}^1 + \alpha_2\lambda_2\mathbf{x}^2 + \ldots + \alpha_k\lambda_k\mathbf{x}^k = 0$$

Multiplying this relation by λ_k and subtracting the product from the original expression, we get

$$\alpha_1(\lambda_1 - \lambda_k)\,\mathbf{x}^1 + \alpha_2(\lambda_2 - \lambda_k)\,\mathbf{x}^2 + \ldots + \alpha_{k-1}(\lambda_{k-1} - \lambda_k)\,\mathbf{x}^{k-1} = 0$$

Since the eigenvalues $\lambda_1, \ldots, \lambda_k$ are distinct, we have $\alpha_1 = \alpha_2 = \ldots = \alpha_{k-1} = 0$, which contradicts our assumption on the linear independence of the vectors $\mathbf{x}^1, \ldots, \mathbf{x}^{k-1}$.

From this theorem it follows that if all the eigenvalues of $\mathbf{A}$ are distinct, then all n its eigenvectors will be linearly independent. In this case the matrix is diagonalizable.

Consider a matrix $\mathbf{V}$ defined by (4.93) whose columns are the eigen-

vectors of A and which is related to A by Eq. (4.92). If all eigenvalues of A are distinct, then its eigenvectors are linearly independent. From the property (5) of determinants (see Sect. 4.6.2) it follows that the matrix V is then nonsingular and the inverse $\mathbf{V}^{-1}$ exists. Premultiplying both sides of (4.92) by $\mathbf{V}^{-1}$, we obtain

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{V}^{-1}\mathbf{V}\Lambda \text{ or } \mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \Lambda \tag{4.96}$$

Thus, the *diagonalizing matrix* is the matrix V constituted by the eigenvectors of A.

4.7 Symmetric Matrices and Quadratic Forms

4.7.1 Eigenvectors and Eigenvalues of Real Symmetric Matrices

A real square matrix A is said to be *symmetric* if all its elements satisfy the relation $a_{ij} = a_{ji}$ for all i, j, i.e. if this matrix is equal to its transpose so that $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$.

Symmetric matrices possess two properties that are significant for many applications, namely,

(i) the eigenvalues of a real symmetric matrix are *real-valued*,

(ii) the eigenvectors corresponding to distinct eigenvalues of a real symmetric matrix are *orthogonal*.

We demonstrate these properties for a 2 by 2 matrix. The eigenvalues of this matrix can be found from the equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{12} & a_{22} - \lambda \end{vmatrix} = 0$$

or

$$\lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}^2 = 0$$

whence

$$\lambda_{1,2} = \frac{1}{2}[a_{11} + a_{22} \pm \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}^2)}]$$

Since the radicand

$$(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}^2) = (a_{11} - a_{22})^2 + 4a_{12}^2$$

is non-negative, λ_1 and λ_2 are always real values.

Denote by

$$\mathbf{x}^1 = \begin{pmatrix} x_1^1 \\ x_2^1 \end{pmatrix} \text{ and } \mathbf{x}^2 = \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix}$$

the eigenvectors of A corresponding to distinct eigenvalues λ_1 and λ_2. We have learned above that the components of the vectors satisfy the relations

$$x_1^1/x_2^1 = -a_{12}/(a_{11} - \lambda_1) = -(a_{22} - \lambda_1)/a_{12}$$
$$x_1^2/x_2^2 = -a_{12}/(a_{11} - \lambda_2) = -(a_{22} - \lambda_2)/a_{12}$$

These relations may be combined into the ratio

$$x_1^1 x_1^2/(x_2^1 x_2^2) = (a_{22} - \lambda_1)/(a_{11} - \lambda_2)$$

or

$$x_2^1 x_2^2 = (a_{11} - \lambda_2) x_1^1 x_1^2/(a_{22} - \lambda_1)$$

The scalar product of the eigenvectors $(\mathbf{x}^1)^T \mathbf{x}^2$

$$(\mathbf{x}^1)^T \mathbf{x}^2 = x_1^1 x_1^2 + x_2^1 x_2^2 = x_1^1 x_1^2 [1 + (a_{11} - \lambda_2)/(a_{22} - \lambda_1)]$$
$$= x_1^1 x_1^2 [a_{11} + a_{22} - (\lambda_1 + \lambda_2)]/(a_{22} - \lambda_1)$$

In view of $\lambda_1 + \lambda_2 = a_{11} + a_{12}$ this product is zero, so that the vectors $\mathbf{x}^1$ and $\mathbf{x}^2$ are orthogonal.

The above properties of a real symmetric matrix may be readily generalized onto the case of a matrix A of any order. From these properties it follows that if $\lambda_i \neq \lambda_j$, then

$$(\mathbf{x}^i)^T \mathbf{x}^j = 0, \quad j \neq i \tag{4.97}$$

that is, the eigenvectors $\mathbf{x}^i$ and $\mathbf{x}^j$ are orthogonal.

Because the eigenvectors are defined accurate to a constant multiplier, this multiplier can be selected for any eigenvector $\mathbf{x}^i$ so that

$$(x^i)^T \mathbf{x}^i = 1, \; i = 1, \ldots, n \tag{4.98}$$

The eigenvectors satisfying this relation are called *normalized*.

4.7.2 Diagonalization of Symmetric Matrices

Consider a vector V whose columns are the eigenvectors of a symmetric matrix A. The transpose $\mathbf{V}^T$ is then a matrix whose rows are the eigenvectors of A.

The element (i, j) of $\mathbf{V}^T \mathbf{V}$ will be the scalar product of row i of $\mathbf{V}^T$ by column j of V, i.e. the quantity $(\mathbf{x}^i)^T \mathbf{x}^j$. In view of (4.97) and (4.98) $\mathbf{V}^T \mathbf{V} = \mathbf{I}$, so that

$$\mathbf{V}^T = \mathbf{V}^{-1} \tag{4.99}$$

Consequently, a symmetric matrix is diagonalized according to the relation

$$\mathbf{V}^T \mathbf{A} \mathbf{V} = \Lambda \tag{4.100}$$

This conclusion holds only for the case when all the eigenvalues of the matrix A are distinct.

4.7.3 Quadratic Forms

Let $\mathbf{x} = (x^1, \ldots, x^n)^T$ be a vector in an n-dimensional linear space with real-valued components. An expression of the form

$$Q(\mathbf{x}) = \sum_i \sum_j a_{ij} x^{(i)} x^{(j)} \tag{4.101}$$

is called a *quadratic form*. Without loss of generality we may let in this expression $a_{ij} = a_{ji}$. If this is not the case, then after some manipulation

$$a_{ij}x^{(i)}x^{(j)} + a_{ji}x^{(j)}x^{(i)} = (a_{ij} + a_{ji})\, x^{(i)}x^{(j)}$$
$$= \frac{1}{2}(a_{ij} + a_{ji})\, x^{(i)}x^{(j)} + \frac{1}{2}(a_{ij} + a_{ji})\, x^{(j)}x^{(i)}$$

we reduce the quadratic form to a form satisfying the above condition. Therefore the numbers a_{ij} can be viewed as elements of the square matrix $\mathbf{A}$ which is *real* and *symmetric*.

A quadratic form is convenient to write in vector matrix notation. For this purpose we rewrite (4.101) as

$$Q(\mathbf{x}) = \sum_i x^{(i)} \sum_j a_{ij} x^{(j)} = \sum_i x^{(i)} z^{(i)} = \mathbf{x}^T \mathbf{z} \tag{4.102}$$

where

$$z^{(i)} = \sum_j a_{ij} x^{(j)}, \quad i = 1, \ldots, n$$

are the components of vector $\mathbf{z}$ related to vector $\mathbf{x}$ by the linear transformation

$$\mathbf{z} = \mathbf{A}\mathbf{x} \tag{4.103}$$

Thus the quadratic form $Q(\mathbf{x})$ can be written as

$$Q(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x} \tag{4.104}$$

The quadratic form $Q(\mathbf{x})$ and the matrix $\mathbf{A}$ are said to be *positive definite* if $Q(\mathbf{x}) > 0$. If $Q(\mathbf{x}) < 0$, then the quadratic form $Q(\mathbf{x})$ and $\mathbf{A}$ are *negative definite*.

In order to determine the conditions for a matrix $\mathbf{A}$ to be positive (negative) definite we consider the linear transformation $\mathbf{x} = \mathbf{V}\mathbf{y}$, where $\mathbf{V}$ is a non-singular matrix. Then

$$Q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{V}^T\mathbf{A}\mathbf{V}\mathbf{y} = Q(\mathbf{y})$$

where the matrix associated with the quadratic form $Q(\mathbf{y})$ is $\mathbf{V}^T\mathbf{A}\mathbf{V}$. Let $\mathbf{V}$ be the diagonalizing matrix, then $\mathbf{V}^T\mathbf{A}\mathbf{V} = \Lambda$ so that

$$Q(\mathbf{x}) = Q(\mathbf{y}) = \mathbf{y}^T\Lambda\mathbf{y} = \sum_i \lambda_i [y^{(i)}]^2 \tag{4.105}$$

As can be seen a necessary and sufficient condition for a matrix **A** to be positive (negative) definite is that all its eigenvalues λ_i be positive (negative).

4.7.4 Use of Quadratic Forms in Search for Extrema of Functions of Many Variables

We illustrate the use of quadratic forms by considering the problem of extrema of functions that is of value for many applications [36].

Let $\mathbf{u} = (u_1, \ldots, u_n)^T$ be a multidimensional variable, $f(\mathbf{u}) = f(u_1, \ldots, u_n)$ be a function defined over the domain $a_i \leqslant u_i \leqslant b_i$, $i = 1, \ldots, n$, and expandable in a converging Taylor series at every point $\mathbf{c} = (c_1, \ldots, c_n)^T$ within this domain, viz.,

$$f(\mathbf{u}) = f(\mathbf{c}) + \sum_i \frac{\partial f}{\partial c_i} \Delta u_i + \frac{1}{2} \sum_i \sum_j \frac{\partial^2 f}{\partial c_i \partial c_j} \Delta u_i \Delta u_j + \ldots \quad (4.106)$$

where

$$\Delta u_i = u_i - c_i, \quad i = 1, \ldots, n$$

$$\frac{\partial f}{\partial c_i} = \frac{\partial f}{\partial u_i}\Big|_{u_i = c_i}, \quad i = 1, \ldots, n$$

$$\frac{\partial^2 f}{\partial c_i \partial c_j} = \frac{\partial^2 f}{\partial u_i \partial u_j}\Big|_{u_i = c_i,\ u_j = c_j}, \quad i, j = 1, \ldots, n$$

Extrema of such functions are known to be achieved in the domain of definition only as stationary points, i.e. where $df/du_i = 0$ for $i = 1, \ldots, n$. Let $\mathbf{u} = \mathbf{c}$ be a stationary point. Then $df/dc_i = 0$, $i = 1, \ldots, n$ and the expansion of $f(\mathbf{u})$ in a Taylor series at $\mathbf{u} = \mathbf{c}$ takes the form

$$f(\mathbf{u}) = f(\mathbf{c}) + \frac{1}{2} \sum_i \sum_j \frac{\partial^2 f}{\partial c_i \partial c_j} \Delta u_i \Delta u_j + \ldots \quad (4.107)$$

Denoting

$$x_i = \Delta u_i = u_i - c_i \quad (4.108)$$

$$a_{ij} = a_{ji} = \frac{\partial^2 f}{2 \partial c_i \partial c_j} \quad (4.109)$$

we rewrite the series (4.107) as

$$f(\mathbf{u}) = f(\mathbf{c}) + Q(\mathbf{x}) + \ldots \quad (4.110)$$

where

$$Q(\mathbf{x}) = \sum_i \sum_j a_{ij} x_i x_j \quad (4.111)$$

is the quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ with the real symmetric matrix **A**.

Recognizing that at $\mathbf{u} = \mathbf{c}$ $x_i = 0$, $i = 1, \ldots, n$, from (4.110) it follows that the stationary point $\mathbf{u} = \mathbf{c}$ has a relative minimum

if $Q(\mathbf{x}) > 0$, and a relative maximum if $Q(\mathbf{x}) < 0$, i.e.— $Q(\mathbf{x}) > 0$. If near a stationary point $Q(\mathbf{x})$ takes on positive and negative values, then the character of this point deserves a separate investigation.

Thus, evaluation of the nature of a stationary point of a function of many variables reduces to the investigation of whether the matrix A of the square form $Q(\mathbf{x})$ is positive (negative) definite.

4.7.5 Vector Derivatives

In order to investigate stationary points of multivariate functions $f(x_1, \ldots, x_n)$ we have to calculate the derivatives.

$$\partial f(x_1, \ldots, x_n)/\partial x_k, \quad k = 1, \ldots, n \tag{4.112}$$

If we represent the set of variables $x_1, \ldots, x_n$ as a vector $\mathbf{x} = (x_1, \ldots, x_n)^T$, then the set of these derivatives may also be represented as the vector

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \partial f(x_1, \ldots, x_n)/\partial x_1 \\ \cdots\cdots\cdots \\ \partial f(x_1, \ldots, x_n)/\partial x_n \end{pmatrix} \tag{4.113}$$

that is referred to as the *vector derivative* of $f(\mathbf{x})$.

In what follows we shall deal with such scalar functions of many variables as the scalar product $\mathbf{a}^T\mathbf{x} = \mathbf{x}^T\mathbf{a}$ and the square form $Q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$. In view of the definitions of these functions (4.35) and (4.101) and the components of the respective vector derivatives (4.113), the relations defining the vector derivatives of a scalar product and quadratic form are as follows:

$$\partial(\mathbf{a}^T\mathbf{x})/\partial\mathbf{x} = \partial(\mathbf{x}^T\mathbf{a})/\partial\mathbf{x} = \mathbf{a} \tag{4.114}$$

$$\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})/\partial\mathbf{x} = 2\mathbf{A}\mathbf{x} \tag{4.115}$$

4.8 Convex Sets

4.8.1 The Concept of a Hypersphere

It is often convenient to transfer analogs of plane geometric figures and geometric solids onto multidimensional space R^n. The concepts of a hypersphere, hyperplane, and a straight line in R^n are most important.

The *hypersphere* is a closed surface all points of which are separated by the same distance r from a certain fixed point a, i.e. these points obey the equation

$$d(\mathbf{x}, \mathbf{a}) = r \tag{4.116}$$

The set of points satisfying the condition $d(\mathbf{x}, \mathbf{a}) < r$ represents the interior points of the hypersphere and is referred to as an *open sphere*. The set of points obeying the condition $d(\mathbf{x}, \mathbf{a}) \leqslant r$ is called the *closed sphere*.

Example 4.8. The equation of a hypersphere in space R^2 may be written as $d^2(\mathbf{x}, \mathbf{a}) = r^2$. If we define this distance by Eq. (4.2), then the hypersphere becomes a circle $(x^{(1)} - a^{(1)})^2 + (x^{(2)} - a^{(2)})^2 = r^2$.

4.8.2 Bounded and Finite Sets

A set X may be called *bounded* if the distance between any two points of this set is bounded, that is, if there exists a number M such that for any $\mathbf{x}_1$ and $\mathbf{x}_2$ in X the distance between them $d(\mathbf{x}_1, \mathbf{x}_2) \leqslant M$.

It may be shown that the necessary and sufficient condition for the set X to be bounded is its location within a certain hypersphere, i.e. the existence of a point a and a number r such that $d(\mathbf{x}, \mathbf{a}) \leqslant r$ for all $\mathbf{x} \in X$.

A set X is called *finite* if it contains a finite number of points. A finite set is always bounded.

4.8.3 Open and Closed Sets

Division of sets into those called open and closed is closely connected with the concept of the neighborhood of a point. The neighborhood of a point $\mathbf{x}$ in space R^n is a sphere of radius ε centered on $\mathbf{x}$. We shall denote the neighborhood of a point by $O_\varepsilon(\mathbf{x})$.

A point $\mathbf{x}$ is said to be an *interior* point of the set X if there exists a neighborhood $O_\varepsilon(\mathbf{x})$ whose points belong to X.

A point $\mathbf{x}$ is a boundary point of X if any neighborhood $O_\varepsilon(\mathbf{x})$ contains both points belonging to X and those not in X. The set of all boundary points forms the *boundary* of X.

A set X is said to be *open* if all its points are interior. A set X is *closed* if in addition to the interior points it contains all its boundary points.

To provide a geometric insight into these concepts consider a set X given in a two-dimensional space (u, v) by the inequality

$$u^2 + v^2 \leqslant r^2$$

This set is a circle of radius r centered at the origin, as shown in Fig. 4.8.

For any point $\mathbf{x} = (u, v)$ of this set we may write either

$$u^2 + v^2 < r^2$$

or

$$u^2 + v^2 = r^2$$

In the first case we may find such a neighborhood $O_\varepsilon(\mathbf{x})$ of $\mathbf{x} = (u, v)$ that the inequality will hold for any point of this neighborhood, and then $\mathbf{x}$ is an interior point of the set.

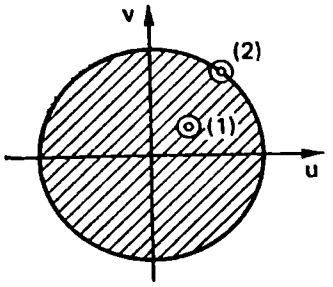


Fig. 4.8. Interior (*1*) and boundary (*2*) points of a set

If the point $\mathbf{x} = (u, v)$ satisfies the last equation, then in any neighborhood $O_\varepsilon(\mathbf{x})$ there exist points $\mathbf{x}_1 = (u_1, v_1)$ and $\mathbf{x}_2 = (u_2, v_2)$ such that $u_1^2 + v_1^2 < r^2$ and $u_2^2 + v_2^2 > r^2$, that is, points that are members of X and points that are not. In this case $\mathbf{x}$ is a boundary point, and the equation defines the boundary of the initial set.

Since the initial set X is specified by a less or equal inequality, the boundary is a subset of the set X. Consequently, the set X is closed. If we use the strict inequality $u^2 + v^2 < r^2$ to specify X, then the boundary will no longer be in X. We obtain an open set.

4.8.4 Hyperplanes and Half Spaces

In linear space R^n hyperplanes are linear subspaces of dimension $n - 1$. For example, in the three-dimensional space R^3, hyperplanes are geometric planes, while in the two-dimensional space R^2 the hyperplanes are just straight lines.

The equation of a hyperplane $L(S)$ may be obtained by specifying in R^n a set S consisting of $n - 1$ linearly independent points $\mathbf{x}_1 \ldots$, $\mathbf{x}_{n-1}$. In agreement with (4.27) any point of a hyperplane

$$\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^{\mathrm{T}} \in L(S)$$

can be written as

$$\mathbf{x} = \sum_{j=1}^{n-1} \alpha_j \mathbf{x}_j \tag{4.117}$$

or, in the coordinate form

$$x^{(i)} = \sum_{j=1}^{n-1} \alpha_j x_j^{(i)}, \quad i = 1, \ldots, n \tag{4.118}$$

Equation (4.117) is not a convenient method for description of a hyperplane, for it involves arbitrary points $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$. On the other hand, it can be demonstrated that these points define a vector $\mathbf{p} = (p_1, \ldots, p_n)$ such that the equation

$$\sum_{i=1}^{n} p_i x_j^{(i)} = 0 \tag{4.119}$$

represents the equation $L(S)$. Substituting the coordinates $x^{(i)}$ from (4.118) into (4.119), we get

$$\sum_{i=1}^{n} p_i \sum_{j=1}^{n-1} \alpha_j x_j^{(i)} = \sum_{j=1}^{n-1} \alpha_j \sum_{i=1}^{n} p_i x_j^{(i)} = 0$$

Because α_j may be any real values, this expression is zero only if

$$\sum_{i=1}^{n} p_i x_j^{(i)} = 0, \quad j = 1, \ldots, n \tag{4.120}$$

The system of homogeneous equations (4.120) enables the components of vector $\mathbf{p}$ to be determined accurate to a constant multiplier, thus proving that Eq. (4.119) is a hyperplane.

The left-hand side of equation (4.119) is the scalar product of vectors $\mathbf{p}$ and $\mathbf{x}$ in linear space R". Therefore, the equation of hyperplane $L(S)$ is usually written as

$$\mathbf{px} = 0 \tag{4.121}$$

This equation defines a hyperplane passing through the origin and at right angles to vector $\mathbf{p}$.

A more general form of the equation of hyperplane results if we shift the origin to an arbitrary point $\mathbf{x}_0$, which corresponds to substituting $\mathbf{x} - \mathbf{x}_0$ for $\mathbf{x}$ in (4.121). Performing this substitution and denoting $\mathbf{px}_0 = c$, we get

$$\mathbf{px} = c \tag{4.122}$$

This is the general equation of hyperplane orthogonal to vector $\mathbf{p}$.

Two sets of points in R^n defined as

$$L^+ = \{\mathbf{x}: \mathbf{px} > c\} \text{ and } L^- = \{\mathbf{x}: \mathbf{px} < c\} \tag{4.123}$$

are called open half spaces defined by a hyperplane L.

4.8.5 Straight Line. Weighted Average

In view of (4.84) at a fixed $\mathbf{x}_1$ and a variable λ the equation

$$\mathbf{x} = \lambda \mathbf{x}_1 \tag{4.124}$$

is a one-dimensional linear space, i.e. a straight line passing through the origin and $\mathbf{x}_1$.

We obtain the equation of a straight line passing through two fixed points $\mathbf{x}_1$ and $\mathbf{x}_2$ from (4.124) by shifting the origin to point $\mathbf{x}_2$,

i.e. by substituting $x - x_2$ and $x_1 - x_2$ for x and x_1, namely,

$$x - x_2 = \lambda (x_1 - x_2)$$

or in a more convenient form

$$x = \lambda x_1 + (1 - \lambda) x_2 \tag{4.125}$$

The values $x = x_1$ and $x = x_2$ are obtained, respectively, by putting $\lambda = 1$ and $\lambda = 0$.

The equation of a line segment connecting points x_1 and x_2 results from (4.125) by constraining the variation of λ by the bounds $0 \leqslant \lambda \leqslant 1$. If we denote $\lambda = \lambda_1$ and $1 - \lambda = \lambda_2$, then the equation of this line segment will be written as

$$x = \lambda_1 x_1 + \lambda_2 x_2;\ \lambda_1,\ \lambda_2 \geqslant 0,\ \lambda_1 + \lambda_2 = 1 \tag{4.126}$$

The quantity x defined by Eq. (4.126) is called the *weighted average* over the elements x_1 and x_2 taken with the weights λ_1 and λ_2. This formula has a simple physical meaning. If we place weights λ_1 and λ_2 at points x_1 and x_2, then x yields the center of gravity of the system.

The concept of weighted average may be expanded to a greater number of points. We define such a weighted average as

$$x = \sum_{i=1}^{m} \lambda_i x_i;\ \lambda_i \geqslant 0, \qquad \sum_{i=1}^{m} \lambda_i = 1 \tag{4.127}$$

We shall call a point x found with Eq. (4.124) a *convex combination* of points x_i. This is a particular case of the linear combination that can be obtained by (4.127), omitting the constraints on the quantities λ_i.

4.8.6 Convex Sets

Let S be a subset of a linear set R^n. The set X is called *convex* if a straight line connecting any two points of this set lies entirely in this set. In other words, S is a convex set if for any $x_1,\ x_2 \in S$ and any $0 \leqslant \lambda \leqslant 1$ we have

$$\lambda x_1 + (1 - \lambda) x_2 \in S$$

Figure 4.9 shows examples of convex (*a* through *d*) and non-convex (*e* through *h*) sets on the plane. Examples of convex sets in a multidimensional space are the entire linear space R^n, its any linear subspace (straight line, hyperplane), half space, convex polyhedrons which are convex sets with linear boundaries (straight lines, hyperplanes). In general, a set that can be expressed as the intersection of a finite number of half spaces is called a *polytope*.

A hypersphere is not a convex set. However, an open sphere and closed sphere are convex sets.

Theorem 4.2. The intersection of convex sets yields a convex set.
Proof. Let **p** and **q** be arbitrary points belonging to the intersection

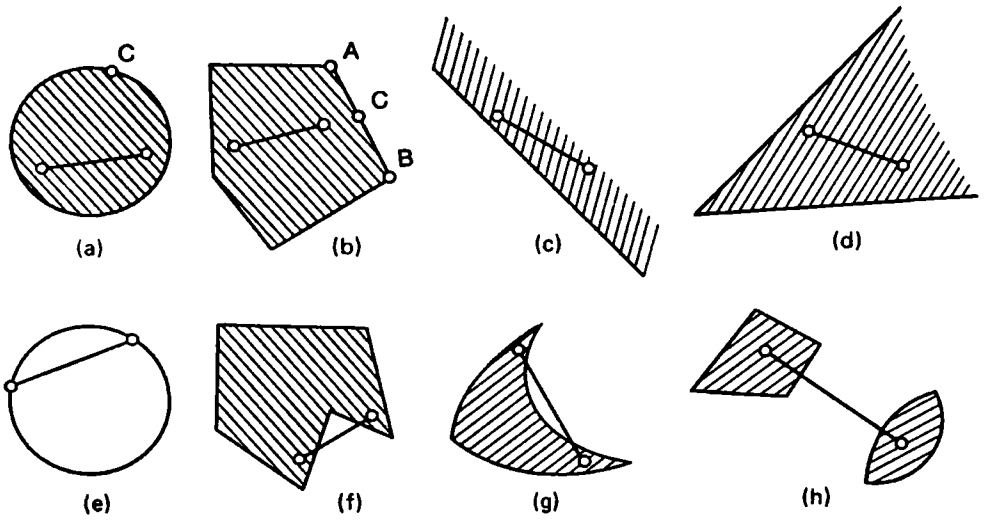


Fig. 4.9. Convex and non-convex sets

of convex sets X and Y. The straight line connecting these points will belong to both X and Y, hence to their intersection.

4.8.7 Convex Hull of a Finite Set

Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ be a finite set of points in space R^n. This finite set is not convex. However, points of A may be elements of convex sets, say S_1, S_2, etc. Then A is a subset of the intersection of convex sets, $S_1 \cap S_2 \cap \ldots$.

The *convex hull* of the set A is the intersection of all convex sets having A as their subset. Specifically, if $A \subset S_1$ and $A \subset S_2$, then

$$\text{co}\,(A) \subset S_1 \cap S_2$$

From this definition it follows that the convex hull co (A) is the smallest convex set containing A. Indeed, if there existed a convex set S such that $A \subset S$ and $S \subseteq \text{co}\,(A)$, then we would have $\text{co}\,(A) \subseteq S \cap \text{co}\,(A)$, that is, $\text{co}\,(A) \subseteq S$. Hence, $S = \text{co}\,(A)$.

Theorem 4.3. The convex hull of a finite set A is a weighted average over the elements of A.

Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ be a finite set and λ_i $(i = 1, \ldots, n)$ be real numbers such that

$$\lambda_i \geqslant 0, \quad \sum_i \lambda_i = 1 \tag{4.128}$$

The quantity

$$\mathbf{x} = \sum_i \lambda_i \mathbf{a}_i \tag{4.129}$$

is called the weighted average over the elements of A. The set S whose members are the quantities x defined by (4.129) for any possible λ_i subject to the constraints (4.128) represents a set of weighted averages over the elements of A. We wish to prove that

$$S = \text{co}\,(A) \tag{4.130}$$

Proof. In order that a set S be the convex hull of the set A, it must be (i) convex and (ii) the least convex set containing A. We prove that both these conditions are met.

(i) Consider two arbitrary systems of weights λ_i and λ_i' defining two points

$$\mathbf{x} = \sum_i \lambda_i \mathbf{a}_i \text{ and } \mathbf{x}' = \sum_i \lambda_i' \mathbf{a}_i$$

of the set S. Take an arbitrary point $\mathbf{x}''$ in the line connecting point x and x',

$$\mathbf{x}'' = \lambda \mathbf{x} + (1-\lambda)\,\mathbf{x}' = \sum_i [\lambda\lambda_i + (1-\lambda)\,\lambda_i']\,a_i = \sum_i \lambda_i'' a_i$$

where $\lambda_i'' = \lambda\lambda_i + (1-\lambda)\,\lambda_i'$.

The quantity λ_i'' is non-negative as it is the weighted average of two non-negative numbers λ_i and λ_i'. In addition,

$$\sum_i \lambda_i'' = \lambda \sum_i \lambda_i + (1-\lambda) \sum_i \lambda_i' = 1$$

Thus $\mathbf{x}''$ is the weighted average over the elements of A, i.e. $\mathbf{x}'' \in S$. Consequently, S is a convex set.

(ii) Let T be an arbitrary convex set containing A. We demonstrate that it contains S as well.

By definition all elements $\mathbf{a}_i$ are members of T. The element

$$\mathbf{x}_2 = \lambda_1 \mathbf{a}_1/(\lambda_1 + \lambda_2) + \lambda_2 \mathbf{a}_2/(\lambda_1 + \lambda_2)$$

is the weighted average of $\mathbf{a}_1$ and $\mathbf{a}_2 \in T$, so that $\mathbf{x}_2 \in T$. On the same lines we have

$$\begin{aligned}\mathbf{x}_3 &= (\lambda_1 + \lambda_2)\,\mathbf{x}_2/(\lambda_1 + \lambda_2 + \lambda_3) + \lambda_3 \mathbf{a}_3/(\lambda_1 + \lambda_2 + \lambda_3) \\ &= \lambda_1 \mathbf{a}_1/(\lambda_1 + \lambda_2 + \lambda_3) + \lambda_2 \mathbf{a}_2/(\lambda_1 + \lambda_2 + \lambda_3) \\ &\quad + \lambda_3 \mathbf{a}_3/(\lambda_1 + \lambda_2 + \lambda_3)\end{aligned}$$

which is the weighted average of $\mathbf{x}_2$ and $\mathbf{a}_3 \in T$, so that $\mathbf{x}_3 \in T$. Continuing this sequence of points, we find that the point

$$\mathbf{x}_m = \frac{\lambda_1 + \ldots + \lambda_{m-1}}{\lambda_1 + \ldots + \lambda_m}\,\mathbf{x}_{m-1} + \frac{\lambda_m \mathbf{a}_m}{\lambda_1 + \ldots + \lambda_m} = \sum \lambda_i \mathbf{a}_i \in S$$

also belongs to the set T. Thus, any $\mathbf{x} \in S$ is also a member of T, i.e. $S \subseteq T$. However, T is an arbitrary convex set containing A. Consequently, S is the least convex set containing A.

The concept of a convex hull may be easily expanded to an arbitrary set. If X is a set of points in R^n (not necessarily convex), then the convex hull of X is the smallest convex set containing X.

4.8.8 Separating and Supporting Hyperplanes

The concepts of separating and supporting hyperplanes are fundamental to linear programming and game theory that will be discussed in subsequent chapters.

Figure 4.10 gives geometric interpretations of these concepts.

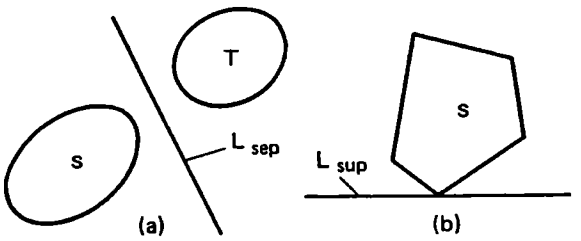


Fig. 4.10. (*a*) Separating and (*b*) supporting hyperplanes

Let S and T be convex nonintersecting sets. The theory of convex sets proves that there exists a hyperplane L_{sep}, called a *separating hyperplane*, such that the sets S and T lie in different half spaces defined by this hyperplane.

If S is a convex set, then there exists a *supporting hyperplane* L_{sup} such that S lies in one of half spaces defined by this hyperplane and at least one point of S belongs to this hyperplane.

A point $\mathbf{x}^*$ in a convex set S is said to be an *extreme point* if there are no two distinct points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S such that $\lambda \mathbf{x}_1 + (1 - \lambda)\, \mathbf{x}_2 = \mathbf{x}^*$ for $0 < \lambda < 1$.

In this definition, λ cannot be either 0 or 1, which implies that the extreme point cannot lie inside a line segment joining any two points of a convex set, but can only be an end point of a line segment.

Clearly, an extreme point should be a boundary point of the set, even though, not all boundary points are extreme points. In the convex polygon in Fig. 4.9*b*, points A and B are extreme points, whereas C is not. If the boundary is curvilinear, as is the case in Fig. 4.9*a*, every boundary point C is extreme. The concept of an extreme point is basis to the definition of a convex polyhedron. A *convex polyhedron* is a convex set having a finite number of extreme points that are the vertices of this polyhedron.

We quote without proof two theorems concerned with extreme points.

Theorem 4.4. Every supporting hyperplane of a convex set contains its extreme point.

Theorem 4.5. The convex set S is a set of weighted averages of its extreme points.

From the last theorem it follows that if S is a convex polyhedron, then any point of this polyhedron is a weighted average of its vertices. Comparing this statement with the definition of a convex hull of a finite set A, we conclude that the convex hull of a finite set A is a convex polyhedron whose vertices are members of A.

EXERCISES

4.1. For three points of three-dimensional space $\mathbf{x} = (1, 0, 2)$, $\mathbf{y} = (3, 4, 0)$ and $\mathbf{z} = (1, 2, 3)$, verify the triangle inequality in spaces $C_2^{(3)}$, $C_1^{(3)}$, and $C^{(3)}$.

4.2. Draw in the x, y plane a circle of radius 1 centered at 0 in spaces $C_2^{(3)}$, $C_1^{(3)}$, and $C^{(3)}$.

4.3. Prove that an open sphere in R^n is a convex set.

4.4. Decide which of the following sets $\{0, 3\}$, $\{5, 7\}$, $\{0, 3\} \cup \{5, 7\}$ are convex in the space R of real numbers.

5

PROBABILITY THEORY AND MATHEMATICAL STATISTICS

5.1 Probability Defined

5.1.1 Events and Sample Space

Theory of probability is underlain by the concept of event which is associated with considering an experiment. In an experiment, an *event* is a result that may or may not occur.

Suppose that under given circumstances the results of the experiment form a set of outcomes

$$Z = \{z_1, \ldots, z_m\} \tag{5.1}$$

We shall refer to the set Z as the *sample description space* of the experiment, and to the elements z of Z as *elementary events* in the sample space Z.

For situations of practical value, we shall be more interested in some collections of events, being subsets of Z, rather than elementary events proper. In the sample space of an experiment, we say an *event* of any subset S of Z,

$$S \subseteq Z \tag{5.2}$$

When we say that an event S has occurred, the implication is that the outcome of the experiment is an elementary event z belonging to S.

For any events S_1 and S_2 belonging to the sample space Z of an experiment, the following definitions take place.

A union of events S_1 and S_2, denoted $S_1 \cup S_2$, is an event when at least one of these events occurs.

We denote by $S_1 \cap S_2$ the *intersection* of events, i.e., an event when both S_1 and S_2 occur. Two events S_1 and S_2 are *disjoint* if the occurrence of one of them precludes the occurrence of the other, i.e. when these events have no elements in common, $S_1 \cap S_2 = \varnothing$.

The complements $\bar{S}$ of an event S is an event indicating that S has not occurred.

A certain event implies that at least one of events of space Z has occurred.

An impossible event signifies the empty set $\varnothing$ implying that none of events of space Z has occurred.

Example 5.1. Two ordinary dice are thrown. Here the possibility set of numbers which turn up is $J = \{1, 2, \ldots, 6\}$. The outcome of each experiment will be an ordered pair $z = \{i, k\}$, where i is the number that turns up in throwing the first dice, and k is the number on the second dice, both k and i belong to J. The outcomes of these experiments are conveniently represented by points on the (i, k) plane as shown in Fig. 5.1. As can be seen the space of experiment consists of 36 points, namely, $Z = \{(1, 1), (1, 2), \ldots, (6, 6)\}$.

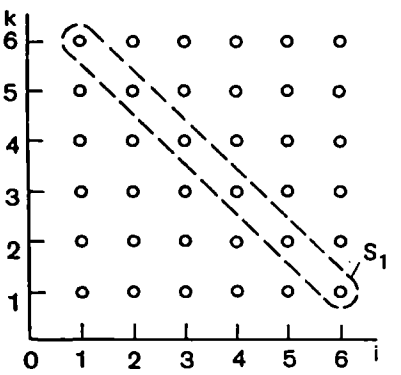


Fig. 5.1. Sample space for two dice showing the set of points defined by S_7

Denote by S_j an event of $i + k = j$, then $S_2 = \{(1, 1)\}$, $S_7 = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, $S_{11} = \{(5, 6), (6, 5)\}$, and $S_{12} = \{6, 6\}$.

5.1.2 The Concept of Probability

Probability theory considers every outcome of an experiment $z \in Z$ as a random event. This means that we do not know in advance which outcome occurs. It is clear, however, that first of all each experiment ends up with some outcome $z \in Z$ and, secondly, one experiment cannot give two outcomes.

Intuitively the *probability* of an outcome is uderstood as a numerical measure assigned to the objective possibility of the particular outcome. This implies that to each element $z \in Z$ we assign some weight p, i.e. a positive real number obeying certain constraints, namely,

(1) the weight p is higher the more we are convinced that this particular outcome will take place,

(2) the sum of weights p for all possible outcomes of the experiment must be equal to unity.

When we assign to an element $z \in Z$ a weight of $p = 0$, we state that this outcome is impossible, i.e. is an *impossible event*. Conversely, if we assign $p = 1$ to an element z in Z, then we are positively sure that this outcome will be a *certain event*. Indeed, in view of the second constraint, all the other elements should be assigned weights $p = 0$, i.e. the events corresponding to these elements are impossible. However, since one outcome has to take place, this will be the one of weight $p = 1$.

The collection of weights assigned to individual elements of the set Z is a set P with elements p and the process of assignment of weights is a mapping of Z onto P. In other words this process defines

probabilities $p \in P$ as functions of outcome $z \in Z$ of experiment, which may be written as $p = p(z)$.

According to the above constraints, the quantities $p(z)$ are real numbers satisfying the conditions

$$p(z) \geqslant 0, \quad \sum_{z \in Z} p(z) = 1 \tag{5.3}$$

The collection of weights $p(z)$ for all $z \in Z$ is called a *probability distribution* over the sample space Z and each weight $p \in P$ assigned to an elementary outcome $z \in Z$ is called the *probability* of this outcome.

If the set Z of elementary outcomes is a set of values of some (not necessarily numerical) variable z, then this variable is called a *random variable*. The collection of weights $p(z)$ represents then a distribution of probabilities of this random variate.

5.1.3 Probability of an Event

We defined events S as some subsets of the set Z of outcomes Z of an experiment. Consequently, an event is a set that consists of a part of elements of Z. Now the probability of an event S, denoted by $P(S)$ or $\mathrm{pr}(S)$, is understood as a sum of weights of the elements constituting this event

$$P(S) = \sum_{z \in S} p(z) \tag{5.4}$$

Two particular cases of this formula are worth noting.

(1) S does not contain elements of Z, i.e. is an empty set. According to (5.4) the probability of an empty set is zero.

(2) S coincides with Z, i.e. contains all generic outcomes of an experiment. Because there is no outcome of the experiment that does not belong to Z, all the space of Z may be considered as a universe of discourse and assigned a weight of $p = 1$. This is in agreement with (5.4) which in view of (5.3) yields

$$P(Z) = \sum_{z \in Z} p(z) = 1 \tag{5.5}$$

5.1.4 Assigning Probabilities to Outcomes

Weights may be assigned to elements z of space Z in a number of ways. Which way to use in a particular situation depends on the nature of the experiment and on the available information. Below we outline four principal methods.

Probability defined as a frequency. Consider an experiment with sample space $Z = \{z_1, \ldots, z_m\}$. This experiment cannot be repeated undefinitely under the same conditions. Assume that N trials have

been carried out and the outcome z of interest has occurred N_z times, i.e. the frequency of this outcome

$$q(z) = N_z/N \tag{5.6}$$

It is not hard to verify that the frequencies of all outcomes z in Z satisfy the conditions (5.3).

When the number of experiments is small, the frequency is substantially random. For example, in tossing a coin 10 times in a series, the number of heads may be 2 in one series, and 8 in another. However, practical experience indicates that as the number of tossings increases, the frequency of individual outcomes loses its random nature to a large degree and shows a trend of approaching, with some insignificant oscillations, some average value, which may be viewed as the probability of this event. This approach of the frequency to probability in increasing the number of experiments differs, however, from coming to a limit in the mathematical sense. For example, 10 heads in 10 tossings of a coin is by no means an impossible result. Because the result of each throw is independent of the results of previous tossings, it is not impossible that 1000 heads occur in 1000 tossing experiments. However, the probability of this event is so small that it may be deemed practically infeasible.

When the number of experiments increases, the frequency approaches the probability in the sense that the probability of somewhat significant deviations of the frequency from probability becomes negligibly small. Consequently, if an experiment can be repeated many times under identical conditions, then frequencies of individual outcomes $q(z)$ may be taken as the probabilities of these outcomes. The question of what number of experiments may be deemed significant in defining a probability by the frequency and what will be the level of significance of such results will be discussed in detail in the chapters devoted to the methods of mathematical statistics.

As a rule, the difficulties associated with multiple replication of an experiment in actual conditions prompt the experimenter to choose other ways of assigning probabilities.

Principle of equirandomization or probable events. This principle is used where one outcome cannot be preferred to others. In this situation, possibilities for any outcome to occur are equal and all outcomes are assigned identical probabilities. If the space of outcomes Z consists of m elements, and an event S contains r elements of Z, then the probability of this event is

$$P(S) = r/m \tag{5.7}$$

By way of example, the probability that a king can be drawn from a well-shuffled pack of cards equals 1/13.

Prior probabilities. Prior probabilities are determined by collect-

ing statistical data about the event in question or about a phenomenon for a long period of time. As a rule in different fields of science, technology, or social life statistical data are collected about events or phenomena that are deemed random and avoid rigorous mathematical description, that is, cannot be predicted with sufficient accuracy. Evidence on the occurrence of such events in the past is collected in order to predict their occurrence in the future.

Posterior probabilities. Prior probabilities yield an adequate description of a phenomenon provided that the conditions under which it occurred in the past continue for the interval of interest. This is not always the case. Therefore, it is important to refine the actual conditions that take place at the instant of experiment. This may be done by conducting a separate experiment. The probabilities that are defined on the basis of available statistical data and by a special experiment are called *posterior probabilities.*

Example 5.2. According to the statistical data collected for many years of regular observations, July in Moscow has 80% of sunny days (event S_1) and 20% of rainy days (event S_2). Accordingly, the prior probabilities that July 12 will be sunny or rainy are $P(S_1) = 0.8$ and $P(S_2) = 0.2$, respectively. However, some years can exhibit considerable deviations from these average figures. A refining of the weather forecast for July 12 with the aid of the forecast for this month issued in June may be considered as an experiment (see also Example 5.5).

5.1.5 Probabilities of Compound Events

Let S_1 and S_2 be some events in the sample space Z of an experiment, the probabilities of these events being $P(S_1)$ and $P(S_2)$. A compound event may be, for instance, a union, intersection, or complement of these simple events. To simplify the calculation of probabilities of compound events we consider some relationships that give the probability of a compound event in terms of the probabilities of the elementary or simple events.

1. If $S_1 \cap S_2 = \varnothing$, then by (5.4)

$$P(S_1 \cup S_2) = \sum_{z \in S_1 \cup S_2} P(z) = \sum_{z \in S_1} P(z) + \sum_{z \in S_2} P(z) = P(S_1) + P(S_2) \tag{5.8}$$

This expression allows for some generalization. Let $S_1, \ldots, S_l$ be mutually exclusive (disjoint) events for which $S_i \cap S_j = \varnothing$ for $i \neq j$. Then

$$P\left(\bigcup_{i=1}^{l} S_i\right) = \sum_{i=1}^{l} P(S_i) \tag{5.9}$$

2. From $X \cup \bar{X} = Z$ and $X \cap \bar{X} = \varnothing$ we find $P(X \cup \bar{X}) = P(X) + P(\bar{X}) = 1$, whence

$$P(\bar{X}) = 1 - P(X) \tag{5.10}$$

3. Let $Y \subseteq X$. Consider the sets Y and $X \setminus Y$ for which $Y \cap (X \setminus Y) = \varnothing$ and $Y \cup (X \setminus Y) = X$. Consequently, $P(X) = P[Y \cup (X \setminus Y)] = P(Y) + P(X \setminus Y)$. Since $P(X \setminus Y) \geqslant 0$, then

$$P(X) \geqslant P(Y), \quad P(X \setminus Y) = P(X) - P(Y) \tag{5.11}$$

4. Let X and Y be two arbitrary events. Consider a subset $B = Y \setminus (X \cap Y)$ satisfying the conditions

$$X \cap B = \varnothing \text{ and } X \cup B = X \cup Y$$

Then

$$\begin{aligned} P(X \cup Y) = P(X \cup B) &= P(X) + P(B) \\ &= P(X) + P[Y \setminus (X \cap Y)] \end{aligned} \tag{5.12}$$

However, $X \cap Y \subseteq Y$ and by virtue of (5.11) $P[Y \setminus (X \cap Y)] = P(Y) - P(X \cap Y)$, therefore

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y) \tag{5.13}$$

Example 5.3. A card is drawn from a pack of 52 playing cards. What is the probability that this card will be (1) a heart or the king of clubs? (2) a heart or a king?

Denote by X the set of hearts, by Y the set of kings, and by V the king of clubs, so that $P(X) = 1/4$, $P(Y) = 1/13$, and $P(V) = 1/52$.

1. Determine $P(X \cup V)$. Since $P(X \cap V) = \varnothing$, then

$$P(X \cup V) = P(\bar{X}) + P(V) = 1/4 + 1/52 = 7/26$$

2. Determine $P(X \cup Y)$. Since $X \cap Y \neq \varnothing$, then $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$. However, if $X \cap Y$ is the king of hearts, then $P(X \cap Y) = 1/52$. Hence,

$$P(X \cup Y) = 1/4 + 1/13 - 1/52 = 4/13$$

5.2 Conditional Probabilities

5.2.1 Conditional Probability Defined

Let T be an event in a sample space Z. By definition the probability of this event is

$$P(T) = \sum_{z \in T} p(z) \tag{5.14}$$

Suppose that we know that some other event S has occurred. We would like to know how the fact that S has occurred influences the probability of T. The probability of an event T conditioned by the fact that another event S has occurred is called the *conditional*

probability of T and denoted by $P(T \mid S)$. We thus wish to find the conditional probability $P(T \mid S)$ in terms of the probabilities of elementary events.

To begin with we simplify the problem. We assume that the set T consists of a single element $z \in Z$. In defining the probability $p(z)$ we assume that any $z \in Z$ can occur. If the number of experiments is sufficiently large, then

$$p(z) = \frac{\text{No of experiments with } z \in Z}{\text{Total number of experiments}} = \frac{N_z}{N} \tag{5.15}$$

If we impose the condition that S has occurred, then all outcomes at which S does not occur should be excluded. Hence, in evaluating $p(z \mid S)$ we should substitute N_S, the number of experiments ended up in S, for N, and $N_{z,S}$, the number of experiments in which both z and S have occurred, for N_z. However, in this case

$$\begin{aligned} (z, S) &= z \cap S = z, \quad \text{if } z \in S \\ &= \varnothing, \quad \text{if } z \notin S \end{aligned} \tag{5.16}$$

Therefore, $N_{z\ S} = N_{z \cap S}$, hence

$$p(z \mid S) = N_{z \cap S}/N_S \tag{5.17}$$

Dividing the numerator and denominator of this expression by the total number of experiments N yields

$$p(z \mid S) = p(z \cap S)/P(S) \tag{5.18}$$

Now we consider the situation when T is an arbitrary subset of Z. Recalling that the probability of a compound event equals the sum of the probabilities of elements constituting this event, we find

$$P(T \mid S) = \sum_{z \in T} p(z \mid S) = \sum_{z \in T} p(z \cap S)/P(S) \tag{5.19}$$

Consider the sum in the numerator of this expression. It consists solely of addends for $z \in T$, for $p(z \cap S) = 0$. Consequently, this sum is constituted by the addends for which $z \in T$ and $z \in S$. Hence, addition may be carried out in $z \in T \cap S$. However, for $z \in S$ we have $p(z \cap S) = p(z)$. Hence,

$$\sum_{z \in T} p(z \cap S) = \sum_{z \in T \cap S} p(z) = P(T \cap S) \tag{5.20}$$

Now,

$$P(T \mid S) = P(T \cap S)/P(S) \tag{5.21}$$

which is often written as

$$P(T \cap S) = P(T \mid S)\, P(S) \tag{5.22}$$

In other situations, that are also frequent, our knowledge of the event S having occurred does not influence the probability of event T.

In other words, the probability of T is independent of whether or not S has occurred, or

$$P(T \mid S) = P(T) \tag{5.23}$$

Such events T and S are said to be *independent* and the formula

$$P(T \cap S) = P(T)\, P(S) \tag{5.24}$$

is sometimes called the *factorization rule* for $P(T \cap S)$.

Example 5.4. A coin is tossed three times. Let T be an event of the coin coming up with one side, and S be an event of the head occurring not more than once. Are these events independent?

Of eight possible outcomes of the experiment, two outcomes correspond to T, four outcomes to S, and one outcome to $T \cap S$. Accordingly, $P(T) = 1/4$, $P(S) = 1/2$, and $P(T \cap S) = 1/8$. Comparing these data with the condition (5.24) we see that T and S are independent.

5.2.2 Conditional Probability Involving More than Two Events

Let an event T occur along with one of mutually exclusive events $S_1, \ldots, S_l$, i.e. disjoint events for which $S_i \cap S_j = \varnothing$ for $i \neq j$. Then

$$T = T \cap (S_1 \cup S_2 \cup \ldots \cup S_l) = (T \cap S_1) \cup (T \cap S_2) \cup \ldots \cup (T \cap S_l)$$

Because S_i are mutually exclusive, so are the events $T \cap S_i$. Consequently, according to (5.9)

$$P(T) = \sum_{i=1}^{l} P(T \cap S_i)$$

or in view of (5.22)

$$P(T) = \sum_{i=1}^{l} p(T \mid S_i)\, P(S_i) \tag{5.25}$$

5.2.3 Bayes' Theorem

Suppose $S_1, \ldots, S_l$ are events for which the prior probabilities $P(S_i)$ are known from a previous experiment. These probabilities seem to be not reliable and to refine them we conduct another experiment.

Let this experiment have resulted in an event T. The conditional probabilities $P(T \mid S_i)$ of T given every mutually exclusive S_i, $i = 1, \ldots, l$ are also known from a previous experiment. Let us look at how the probabilities of $S_1, \ldots, S_l$ change after the result T

of the experiment has been obtained. These probabilities are denoted by $P(S_i \mid T)$, $i = 1, \ldots, l$, and called the posterior of *inverse probabilities* of events $S_1, \ldots, S_l$.

To compute the inverse probabilities we resort to the formula (5.22) rewritten as

$$P(T \cap S_i) = P(T \mid S_i)\, P(S_i) = P(S_i \mid T)\, P(T)$$

whence

$$P(S_i \mid T) = P(T \mid S_i)\, P(S_i)/P(T) \tag{5.26}$$

By virtue of (5.25) the probability $P(T)$ is

$$P(T) = \sum_{i=1}^{l} P(T \mid S_i)\, P(S_i) \tag{5.27}$$

Making use of this formula in (5.26), we obtain an expression known as *Bayes' formula*

$$P(S_i \mid T) = \frac{P(T \mid S_i)\, P(S_i)}{\sum_{i=1}^{l} P(T \mid S_i)\, P(S_i)} \tag{5.28}$$

Example 5.5. The weather forecast for July 12 in Example 5.2 has predicted a sunny day (T_1). This forecast may be in error with the conditional probabilities $P(T_1 \mid S_1) = 0.9$ and $P(T_1 \mid S_2) = 0.3$. The posterior probabilities of S_1 and S_2 will be $P(S \mid T_1) = 0.9 \times 0.8/(0.9 \times 0.8 + 0.3 \times 0.2) = 0.923$; hence, $P(S \mid T_1) = 0.0769$

5.3 Continuous Random Variables and Their Probability Distribution Functions

5.3.1 The Concept of a Continuous Random Variable

Thus far we have assumed that the sample space Z of experiment is a finite set. However in many a situation the outcome z of an experiment may be any value in a certain interval $a \leqslant z \leqslant b$. In other words, z is a *continuous random variable*, for which the sample space is the entire range of its values,

$$Z = \{z:\ a \leqslant z \leqslant b\} \tag{5.29}$$

containing an infinite set of points.

By way of illustration, random variables are heights of trees in a forest, current or voltage in an electric circuit, and ambient temperature at different time instants during the day. Now, we can no longer speak of the probability of one outcome, since with an infinitely large number of possible outcomes the weight of each outcome is $p = 0$. Therefore the distribution of probabilities for a continuous random variable is defined in a manner different from that for

a discrete variable. The key functions of this definition are the *probability density function* (p.d.f.) and *cumulative distribution function* (c.d.f.) also called the probability distribution function (PDF).

5.3.2 Cumulative Distribution Function

Let Z be the sample space of an experiment on a one-dimensional random variable that takes on real values, and R, be the set of all real numbers whose elements will be denoted by x. The cumulative distribution function $F(x)$ defines the probability that the value of z will not exceed a given x,

$$F(x) = P(-\infty < z \leqslant x) \quad (5.30)$$

The general behavior of $F(x)$ may be perceived by recalling some of its properties.

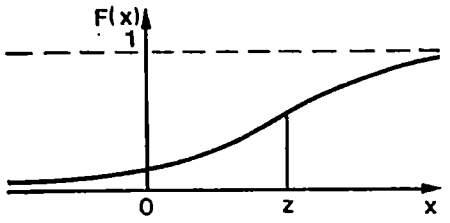


Fig. 5.2. Cumulative distribution function of a continuous random variable

(1) Since for $x_2 > x_1$ the interval $[-\infty, x_2]$ lies within the interval $[-\infty, x_1]$, then according to (5.30) $F(x_2) \geqslant F(x_1)$. Thus, $F(x)$ is a non-decreasing function.

(2) $F(-\infty) = 0$, because the outcome of the experiment $z \in (-\infty, x]$ becomes an impossible event as x tends to $-\infty$.

(3) $F(+\infty) = 1$, since the outcome of the experiment $z \in (-\infty, x]$ becomes a true event as x tends to infinity.

A function satisfying these conditions has a form shown in Fig. 5.2.

5.3.3 Probability Density Function

If the cumulative distribution function is known, we can calculate the probability that the value of the random variable will be within a small interval from x to $x + \Delta x$. According to (5.30) this probability is

$$F(x + \Delta x) - F(x) = [F(x + \Delta x) - F(x)]\, \Delta x/\Delta x \quad (5.31)$$

The first cofactor on the right-hand side of this expression is the probability per unit length of Δx. If the limit of this ratio exists for $\Delta x \to 0$, which is usually the case for most of continuous random variables, the resulting function is called the *probability density function* (p.d. f.) or simply *probability density* and denoted by $w(x)$ or $f(x)$. Thus,

$$w(x) = \lim_{\Delta x \to 0} \{[F(x + \Delta x) - F(x)]/\Delta x\} = \mathrm{d}F(x)/\mathrm{d}x \quad (5.32)$$

that is

$$F(x + \Delta x) - F(x) = w(x)\,\Delta x \tag{5.33}$$

The general behavior of $w(x)$ is illustrated by the graph in Fig. 5.3.

Let us look at some properties of this function. Suppose that a and b are arbitrary points on the real axis, $b > a$. Consider the integral

$$\int_a^b w(x)\,dx$$

By formally replacing $w(x)\,dx$ with $dF(x)$, we get

$$\int_a^b w(x)\,dx = \int_{F(a)}^{F(b)} dF(x) = F(b) - F(a) \tag{5.34}$$

However, according to (5.30) this expression gives the probability

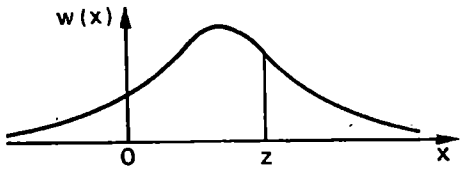


Fig. 5.3. Probability density function

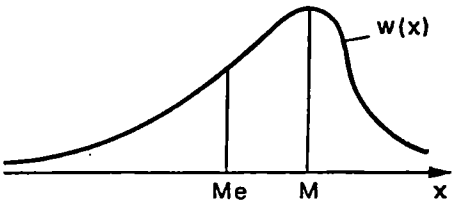


Fig. 5.4. Relationship between the mode and median for a skewed distribution

of finding the random variable z in the interval (a, b). Thus,

$$P(a < z < b) = \int_a^b w(x)\,dx \tag{5.35}$$

As a particular case, from (5.35) it follows

$$\int_{-\infty}^{\infty} w(x)\,dx = 1 \tag{5.36}$$

This expression indicates that the area under the curve of probability density $w(x)$ is always unity. It is called the *normalization condition.*

Comparing (5.30) and (5.35), we find the relation between $F(x)$ and $w(x)$ in the integral form

$$F(x) = \int_{-\infty}^{x} w(\xi)\,d\xi \tag{5.37}$$

For complete description of the p.d.f. curve we point also to two characteristic values of x called the *mode* or *modal value* and the *median*.

The mode, M, of a random variable is the value of x at which $w(x)$ takes its greatest value, i.e. $w(x) = \max$.

The median, Me, of a random variable is the value $x_{1/2}$ for which $P(z < Me) = P(z > Me)$, i.e. the areas under the probability density curve on the right and on the left of this point are equal.

Figure 5.4 shows the positions of mode and median for a probability density function asymmetric about its maximum value.

The concepts of cumulative distribution function and probability density function can be readily generalized over the case of multidimensional random variable (random vector) [20, 22]. We deter consideration of these topics until Sect. 5.5 and for the moment confine ourselves to discussing some laws of distribution of continuous random variates.

5.3.4 The Continuous Uniform Distribution

If a random variable takes any value within a given interval from α to β with equal probability and cannot take values outside this interval, it is said to be uniformly distributed in the interval $[\alpha, \beta]$ with probability density

$$\begin{aligned} w(z) &= 1/(\beta - \alpha), \quad \alpha \leqslant z \leqslant \beta \\ &= 0, \qquad\qquad \text{otherwise} \end{aligned} \tag{5.38}$$

It is convenient to describe a uniform distribution by the *mean value* $\mu = (\alpha + \beta)/2$ and *spread* $\omega = \beta - \alpha$, and denote it by $R(\mu, \omega)$. In this notation the probability density function becomes

$$\begin{aligned} w(z) &= 1/\omega, \quad \mu - \omega/2 \leqslant z \leqslant \mu + \omega/2 \\ &= 0, \quad z < \mu - \omega/2, \; z > \mu + \omega/2 \end{aligned} \tag{5.39}$$

Example 5.6. In rounding off a number to integer value, errors are equally likely to take on values from -0.5 to 0.5, that is, obey a distribution $R(0, 1)$.

Example 5.7. In measuring by a scale with the value of a division δ, the errors of reading will be random variables distributed as $R(0, \delta)$.

Example 5.8. If buses arrive at 10-min intervals, the waiting time for a passenger unaware of the schedule will be a random variable with the distribution $R(5, 10)$.

5.3.5 The Normal (Gaussian) Distribution

This distribution law is most important of all because it is most frequent in practical applications and is the limiting law by which some other laws can be approximated under fairly frequent typical conditions.

The random variable z is said to have normal distribution if its probability density function is

$$w(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(z-\mu)^2/2\sigma^2}, \quad -\infty < z < \infty \tag{5.40}$$

The parameters μ and σ^2 in this definition are the *mathematical expectation* (*mean*) and *variance*. In what follows we denote a normal distribution by $N(\mu, \sigma^2)$. Figure 5.5 shows the plots of (5.40) for

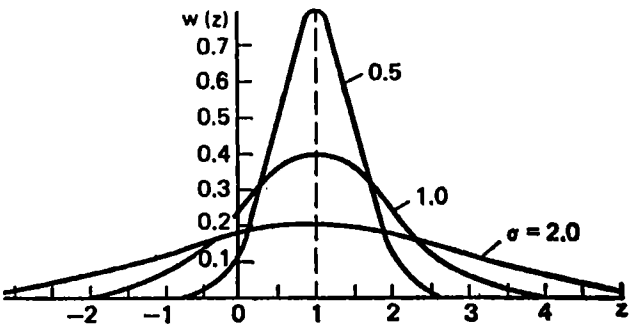


Fig. 5.5. Normal distribution functions for three different standard deviations

$\mu = 1.0$ and three values of σ. It can be seen that as $\sigma \to 0$ the normal probability density $w(z)$ becomes a delta function $\delta(z - \mu)$ (see Sect. 2.5).

An important problem of probability theory is to determine the probability of a random variable lying in the given interval (a, b) on the real axis. For a normal distribution

$$P(a < z < b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-(z-\mu)^2/2\sigma^2}\, dz \tag{5.41}$$

This expression takes on a more convenient form upon the substitution of the *standardized variable* $(z - \mu)/\sigma = t$. Then, $dt = dz/\sigma$ and

$$P(a < z < b) = \frac{1}{\sqrt{2\pi}} \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} e^{-t^2/2}\, dt \tag{5.42}$$

Unfortunately, the integral of this expression cannot be obtained in terms of elementary functions. Therefore, it is tabulated [21] as values of the special function called the *standard normal integral*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{5.43}$$

Now the probability of a random variable lying in the interval (a, b) is

$$P(a < z < b) = \Phi((b - \mu)/\sigma) - \Phi((a - \mu)/\sigma) \tag{5.44}$$

In calculating with this expression it is useful to remember that

$$\Phi(0) = 1/2, \quad \Phi(\infty) = 1, \quad \Phi(-x) = 1 - \Phi(x) \tag{5.45}$$

5.4 Expected Values and Moments

5.4.1 Statistics Defined

The probability distribution function or probability density function completely characterizes a random variable. However, in many problems it is hard or even impossible to completely describe the cumulative distribution function.

On the other hand, many problems can be solved by knowing only some parameters characterizing the random variate. The most popular numerical parameters, called *statistics* or *moments* of a random variable, are the mathematical expectation or expected value and variance or mean square deviation. These parameters are readily determined by experimental data and provide a general insight into the behavior of the random variable under study.

5.4.2 Mathematical Expectation

In many problems the random variable is treated as a possible outcome of some experiment and is expressed by a real number. Examples of random variables assuming numerical values are: electric current consumed by a living block, voltage in the circuit at a given time instant, ambient temperature at a time instant, the number of cosmic-ray particles incident on the earth's surface per unit time, the number of points made by a shot, the number of customers in a shop at a given time, and the like.

For all these situations, an important numerical characteristic of random variables is the mathematical expectation.

Suppose that the elements of the sample space $Z = \{z_1, \ldots, z_m\}$ allow numerical estimation. If after N trials the outcome z_1 has occurred r_1 times, . . ., and the outcome z_m has occurred r_m times, so that $r_1 + \ldots + r_m = N$, then the average or mean value of the random variable z is

$$z_{av} = (z_1 r_1 + \ldots + z_m r_m)/N = \sum_{i=1}^{m} z_i (r_i/N) \tag{5.46}$$

where $r_i/N = q_i$ is the frequency of outcome z_i.

If we increase the number of experiments N, the ratio r_i/N will approach the probability p_i of outcome z_i. By replacing r_i/N in (5.46) with the probability p_i we obtain the formula for the mathematical expectation $E(z)$ or $\bar{z}$ of the random variable z. Dropping the indices of the variables, we get

$$E(z) = \bar{z} = \sum_{z \in Z} zp(z) \tag{5.47}$$

As can be seen, the difference of the mathematical expectation of a random variable from its average value has the same nature as the difference of probability from frequency.

The expression (5.47) is useful in evaluating the expected value when the sample space Z is a finite set. If z is a continuous random variable with probability density function $w(z)$, then $p(z)$ in this expression may be treated as the probability that the value of the random variable lies in the interval from z to $z + \Delta z$, i.e. let $p(z) = w(z)\,\Delta z$. If we let it come to the limit as $\Delta z \to 0$, replace Δz with dz, and the sum for the integral, we obtain

$$E(z) = \bar{z} = \int_{-\infty}^{\infty} zw(z)\,dz \tag{5.48}$$

It will be instructive to note some properties of the mathematical expectation important for our further consideration.

1. The expectation of a non-random quantity equals this quantity. Let c be a non-random quantity, i.e. a constant. This constant may be viewed as the result of an experiment for which the sample space Z consists of only one element c, so that $p(c) = 1$. From (5.47) it follows that

$$E(c) = cp(c) = c \tag{5.49}$$

2. A constant multiplier may be factored out outside the expectation symbol,

$$E(cz) = \sum_{z \in Z} czp(z) = c \sum_{z \in Z} zp(z) = cE(z) \tag{5.50}$$

Example 5.9. The mathematical expectation of a uniformly distributed random variable is

$$\bar{x} = \int_{-\infty}^{\infty} xw(x)\,dx = 1/(b-a)\int_{a}^{b} x\,dx = (b+a)/2$$

5.4.3 Expected Value of a Function of Random Variable

The random variable is not a number in many problems of practical value. For example, items produced by a factory may be qualified as standard and rejects. In tossing a coin the outcome of a throwing may be a head or a tail. A playing card drawn at random from a pack has some value and suit. A radar signal carries information about the presence or absence of a target. In all these examples random variables have qualitative rather than quantitative description. These random variables, however, may also be described in quantitative terms.

We consider the random variable as a result of a certain experiment, either performed by an experimenter, or being a natural process. It should be remembered, however, that any experiment requires certain costs to be incurred out and is done not for curiosity but to achieve a certain objective. An outcome of the experiment may be favorable or unfavorable depending on how close it brings us to the objective. When the objective is reached, we may treat the result as a winning or profit while failure to achieve the objective may be treated as a loss. Both profit and loss may be expressed numerically, say as monetary values.

Thus, every outcome of the experiment $z \in Z$ may be associated with some numerical value, i.e. carry out a map f of the sample space Z onto the set of real numbers R, viz.,

$$f : Z \to R \tag{5.51}$$

This map results in a real function $f(z)$ defined on Z, which can be treated as a random variable, specifically, we may determine its expected value.

Let $f(z)$ be some numerical function defined on a set Z. This implies that the elements $\{z_1, \ldots, z_m\}$ of the set Z are associated with the functional values $f(z_1), \ldots, f(z_m)$. The probabilities of these values will be the same as the probabilities of the random variables $z_1, \ldots, z_m$. Therefore, $p(z)$ is a probability distribution for the random variable z and for the function of random variate $f(z)$. Hence, the mathematical expectation of a function of random variable can be obtained by the formulas for expected value of a random variable

by substituting $f(z)$ for z. To be more specific, for a discrete random variable we obtain

$$E[f(z)] = \overline{f(z)} = \sum_{z \in Z} f(z)\, p(z) \tag{5.52}$$

and for a continuous random variable,

$$E[f(z)] = \overline{f(z)} = \int_{-\infty}^{\infty} f(z)\, w(z)\, dz \tag{5.53}$$

Example 5.10. The probability that a manufactured item will not pass quality control equals p. What will be the average profit per item if a is its manufacturing cost and b its selling price?

Denote the average profit by Q and in view of (5.52) we find

$$Q = (b - a)(1 - p) - ap = b(1 - p) - a$$

In many a problem the expected value is viewed as the average payoff in a large number of experiments. It is assumed that the experiment is worth conducting if $\overline{f(z)} \geqslant 0$.

Another interpretation of expected value is associated with a bet or game. Suppose that some event may occur with probability p. One of two players agrees to pay a sum of b units if this event does not occur while the other pays him a sum of a if it occurs. Therefore the first player receives a sum of a with probability p and pays a sum of b with probability $1 - p$. His average profit is

$$Q = ap - b(1 - p)$$

The game will be called *fair* if the average payoff is zero, i.e. on the condition

$$a/b = (1 - p)/p$$

Theorem 5.1 (mean value). Substituting the minimum value $\min f(z)$ for $f(z)$ in (5.52), we obtain the statement of the *mean value theorem*

$$\overline{f(z)} \geqslant \min_z f(z) \sum_{z \in Z} p(z) = \min_z f(z)$$

5.4.4 Moments. Variance. Standard Deviation

Moments are important characteristics of random variables. The quantity

$$\alpha_r' = E(z^r), \quad r = 1, 2, \ldots$$

is called the rth moment (about the origin) of the random variable z, and the quantity

$$\alpha_r = E(z - \mu)^r$$

where $\mu = E(z)$ is the rth *moment about the mean* or the rth *central moment*.

For a discrete random variable, the rth moment about the origin is

$$\alpha_r'(z) = \sum_z z^r p(z) \tag{5.54}$$

and for a continuous random variable,

$$\alpha_r'(z) = \int_{-\infty}^{\infty} z^r w(z)\, dz \tag{5.55}$$

It is clear that the expectation $E(z) = \bar{z}$ is the first moment $\alpha_1'(z)$ which is called the *mean value*. The second moment is called the *mean square value*

$$\alpha_2'(z) = \sum_z z^2 p(z) = E(z^2) = \overline{z^2} \tag{5.56}$$

In Soviet literature, the difference between a random variable and its expectation is called the *central* random variable and denoted by $\overset{\circ}{z} = z - \bar{z}$. If we substitute $\overset{\circ}{z}$ of z in (5.54) and (5.55), we obtain the expressions for the respective *central moments*. It is not hard to verify that the first central moment, i.e. the expectation of a central random variable, is zero, viz.,

$$E(\overset{\circ}{z}) = \int_{-\infty}^{\infty} (z-\bar{z})\, w(z)\, dz = \int_{-\infty}^{\infty} zw(z)\, dz - \bar{z} \int_{-\infty}^{\infty} w(z)\, dz = \bar{z} - \bar{z} = 0$$

An important characteristic of a random variable is the second central moment, called the *variance* of a random variable z and denoted var (z) or σ_z^2. The variance of a discrete random variable is given by

$$\text{var}(z) = \sigma_z^2 = \sum_z (z-\bar{z})^2 p(z) \tag{5.57}$$

and for a continuous random variable,

$$\text{var}(z) = \sigma_z^2 = \int_{-\infty}^{\infty} (z-\bar{z})^2 w(z)\, dz \tag{5.58}$$

Comparing (5.57) and (5.58) with (5.52) and (5.53), we see that variance may be viewed as the expectation of the square of a central random variable

$$\sigma_z^2 = E[(z-\bar{z})^2] \tag{5.59}$$

The variance var (z) is a measure of the variability or dispersion of the values of z about the expected value, i.e. characterizes the scatter of this random variable. The smaller this characteristic,

the closer cluster the values of this random variable around the expectation.

In some situations the variance is not convenient for practical usage as it has the dimension of the square of a random variable. Therefore, to characterize scatter of random variables use is made of the square root of the variance

$$\sigma_z = \sqrt{\text{var}(z)} \tag{5.60}$$

called the *standard deviation.*

We note two important properties of the variance:

(1) the variance of a nonrandom variable (constant) is zero

$$\text{var}(c) = E[(c - c)^2] = E(0) = 0 \tag{5.61}$$

(2) for a real constant c,

$$\text{var}(cz) = E[(cz - c\bar{z})^2] = c^2 \text{ var}(z) \tag{5.62}$$

Some other properties of the variance will be discussed in later sections.

Example 5.11. The variance of a uniformly distributed random variable is

$$\sigma_z^2 = \int_{-\infty}^{\infty} (z - \bar{z})^2 w(z)\, dz = \frac{1}{(b-a)} \int_a^b (z - (b+a)/2)^2\, dz = (b-a)^2/12$$

5.5 Two-Dimensional Random Variables

5.5.1 The Notion of a Two-dimensional Random Variable

In carrying out an experiment we are often interested simultaneously in a few random variables, say two, x and y. For example, in drawing a card of a pack we may be interested in its suit and value; in selecting a tree in a forest, these may be its height and diameter; in testing capacitors, their capacitance and breakdown voltage.

In all these examples we deal with two sets of random variables, $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. However, these sets are not independent, but related in a certain way, namely, the outcome of one experiment z yields two random variables $x \in X$ and $y \in Y$, so that $z = (x, y)$. Thus, the set of outcomes of experiment Z is the direct product of the sets X and Y, viz.,

$$Z = X \times Y \tag{5.63}$$

Consider events and their probabilities that are associated with two-dimensional random variables. We shall illustrate the necessary concepts with the use of a pencil box that contains pencils differing in two properties, color $x \in X = \{\text{R, G}\}$ and hardness $y \in Y =$

{H, B}. There is a total of eight pencils with the sets of properties aligned as follows:

R R R R R G G G
H H H B B H B B

The result of an experiment will be a random choice of one pencil. If we would like that the set of properties of the chosen pencil will be $x = \text{R}$ and $y = \text{B}$, then we should consider the following obvious equal probabilities:

(1) $p(x) = p(\text{R}) = 5/8$, the probability of taking a red pencil (hardness is of no significance);

(2) $p(y) = p(\text{B}) = 1/2$, the probability of taking a pencil of soft (B) grade (color is of no significance).

The probabilities $p(x)$ and $p(y)$ are unconditional probabilities of properties x and y, and give probability distribution of sets X and Y satisfying the conditions (5.3).

(3) $p(x, y) = p(\text{R}, \text{B}) = 1/4$, the probability of taking a pencil that is red and soft (there are two of them). This is the joint probability distribution of two properties x and y given over the space $Z = X \times Y$ with elements $z = (x, y)$ and satisfying the conditions

$$p(x, y) \geqslant 0, \quad \sum_{x, y} p(x, y) = 1 \tag{5.64}$$

(4) $p(x \mid y) = p(\text{R} \mid \text{B}) = 1/2$, the conditional probability in which $y = \text{B}$ implies that the choice is only from soft pencils, the hard pencils being excluded from consideration.

In order to join these probabilities together, denote by x_i some outcome $x \in X$, and by y_i some outcome $y \in Y$, and consider two events in space $Z = X \times Y$, $T_i = \{(x_i, y_1), \ldots, (x_i, y_m)\}$ and $S_j = \{(x_1, y_j), \ldots, (x_n, y_j)\}$.

The event T_i implies that the given outcome $x_i \in X$ has occurred and also some $y \in Y$. Because some $y \in Y$ always takes place in the experiment, the event T_i implies just that the outcome x_i has occurred. Therefore, $T_i = x_i$. Similarly, $S_j = y_j$. It is not hard to see that $T_i \cap S_j = (x_i, y_j)$. By virtue of (5.21) we have

$$p(x_i \mid y_i) = p(x_i, y_j)/p(y_j) \tag{5.65}$$

Since x_i is any $x \in X$, and y_j may be any $y \in Y$, then we may drop the indices and write for any $x \in X$ and $y \in Y$

$$p(x \mid y) = p(x, y)/p(y) \tag{5.66}$$

or

$$p(x, y) = p(x \mid y)\, p(y) \tag{5.67}$$

If the probability of outcome x is independent of what outcome y has occurred, then the outcomes x and y are mutually *independent.* For independent outcomes, we have

$$p(x \mid y) = p(x) \tag{5.68}$$

therefore (5.67) takes the form

$$p(x, y) = p(x)\, p(y) \tag{5.69}$$

We note also that in dealing with a two-dimensional variable, it is immaterial which of the quantities x or y stands first and which second. Permuting x and y in (5.67), we get

$$p(x, y) = p(y \mid x)\, p(x) \tag{5.70}$$

Comparing (5.67) and (5.70) yields

$$p(x, y) = p(x)\, p(y \mid x) = p(y)\, p(x \mid y) \tag{5.71}$$

5.5.2 Probability Distribution Function

The expressions (5.66)-(5.71) combine the probabilities of various events for two-dimensional random variables with finite sample spaces. In practice, however, one often encounters continuous random variables X and Y which may be described by analogy with one-dimensional case, i.e. by joint cumulative distribution and probability density functions.

The joint cumulative distribution function $F(x, y)$ defines the probability that a value of random variable X will not exceed a given value x, and a value of random variable Y will not exceed a given value y, viz.,

$$F(x, y) = P(-\infty < X \leqslant x, \quad -\infty < Y \leqslant y) \tag{5.72}$$

The function $F(x, y)$ is a single-valued, real, non-negative function of x and y with the following properties:

(1) $F(x, y)$ is nondecreasing function of x and y;

(2) $F(-\infty, y) = F(x, -\infty) = 0$;

(3) $F(\infty, \infty) = 1$.

Denote by $F_1(x)$ and $F_2(y)$ the c.d.f. of each of the components of the two-dimensional random variable (X, Y). These functions may be obtained from the joint c.d.f. $F(x, y)$ by assuming that one of the components may take on any value from $-\infty$ to ∞, and called *marginal* c.d.f.s,

$$F_1(x) = F(x, \infty), \quad F_2(y) = F(\infty, y) \tag{5.73}$$

The joint probability density function $w(x, y)$ can be defined similar to its one-dimensional analog as

$$w(x, y) = \frac{\partial^2 F(x, y)}{\partial x\, \partial y} \tag{5.74}$$

The quantity $w(x, y)\,dx\,dy$ gives the probability that the first component X of the two-dimensional variable (X, Y) will lie in the interval from x to $x + dx$, and the second component Y will lie in the interval from y to $y + dy$, i.e.

$$w(x, y)\,dx dy = P(x \leqslant X \leqslant x + dx,\ y \leqslant Y \leqslant y + dy) \quad (5.75)$$

The probability that in the two-dimensional random variable (X, Y) the variable X lies in the interval from a to b and the variable Y in the interval from c to d is

$$P(a \leqslant X \leqslant b,\ c \leqslant Y \leqslant d) = \int_a^b \int_c^d w(x, y)\,dx\,dy \quad (5.76)$$

This expression provides a relation between $w(x, y)$ and $F(x, y)$ in integral form. Comparing (5.72) and (5.76) yields

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} w(\xi, \eta)\,d\xi\,d\eta \quad (5.77)$$

This expression may be used to obtain one-dimensional c.d.f.s for the components of the two-dimensional random variables (X, Y). Indeed, by letting $y = \infty$, we get

$$F_1(x) = F(x, \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} w(\xi, \eta)\,d\xi\,d\eta$$

Here we recall that the one-dimensional c.d.f. $F_1(x)$ is related to the one-dimensional p.d.f. $w(x)$ by the formula (5.37), namely,

$$F_1(x) = \int_{-\infty}^{x} w_1(\xi)\,d\xi$$

Comparing the last two relations yields the marginal p.d.f.s

$$w_1(x) = \int_{-\infty}^{\infty} w(x, y)\,dy \quad (5.78)$$

$$w_2(y) = \int_{-\infty}^{\infty} w(x, y)\,dx \quad (5.79)$$

The functions $w_1(x)$ and $w_2(y)$ define the distribution of probabilities of individual components of the random variable (X, Y). However, they cannot characterize this variable completely and have to be augmented by the conditional distribution laws $w(x \mid y)$ and

$w(y \mid x)$ that define the probability density functions of one random variable at a fixed value of the other.

In order to relate the two-dimensional, joint p.d.f. $w(x, y)$ and the conditional distributions $w(x \mid y)$ and $w(y \mid x)$, we consider the real plane $R \times R$ whose points will be denoted (X, Y). Let us evaluate the probability of a point (X, Y) lying in an elementary rectangle R_d of sides dx and dy built at the point (x, y) as shown in Fig. 5.6.

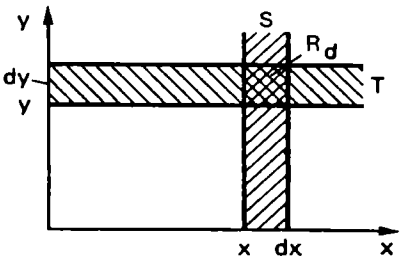


Fig. 5.6. Elementary rectangle R_d in a two-dimensional space

Denote by S and T the events of (X, Y) lying in the interval from x to $x + dx$ and in the interval from y to $y + dy$. The occurrence of this point in the rectangle R_d means the intersection of these events, therefore, by virtue of (5.22) and (5.75) the probability

$$w(x, y)\,dxdy = P(S \cap T) = P(T \| S)\,P(S)$$

Observing that $P(S) = w_1(x)\,dx$ and $P(T \mid S) = w(y \mid x)\,dy$, we have

$$w(x, y)\,dxdy = w(y \mid x)\,dy\,w_1(x)\,dx$$

or

$$w(x, y) = w(y \mid x)\,w_1(x) \tag{5.80}$$

If we fix Y rather than X, this expression rewrites as

$$w(x, y) = w(x \mid y)\,w_2(y) \tag{5.81}$$

These expressions may be used to obtain the conditional p.d.f.s.

$$w(x \mid y) = w(x, y)/w_2(y) \tag{5.82}$$
$$w(y \mid x) = w(x, y)/w_1(x)$$

5.5.3 Expected Values

The expressions we have just derived for the probability density functions allow the notion of expected value to be expanded to the case of a two-dimensional random variable (x, y). From (5.48) it follows that the expectations of the components of the two-dimensional random variable will be

$$\bar{x} = \int_{-\infty}^{\infty} x w_1(x)\,dx, \quad \bar{y} = \int_{-\infty}^{\infty} y w_2(y)\,dy \tag{5.83}$$

The expectation of $f(x, y)$ is

$$\overline{f(x, y)} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, w(x, y)\, dx\, dy \tag{5.84}$$

The conditional expected values $E(x \mid y)$ and $E(y \mid x)$ can be derived with the conditional p.d.f.s $w(x \mid y)$ and $w(y \mid x)$, namely,

$$\begin{aligned} E(x \mid y) &= \int_{-\infty}^{\infty} x w(x \mid y)\, dx \\ E(y \mid x) &= \int_{-\infty}^{\infty} y w(y \mid x)\, dy \end{aligned} \tag{5.85}$$

Using the joint p.d.f. $w(x, y)$, we may easily prove that the *expectation of a sum of random variables equals the sum of the expectations of these variables.* For this purpose we consider $x + y$ as a function of two-dimensional random variable (x, y). By virtue of (5.84) we find

$$\begin{aligned} E(x+y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y)\, w(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} w(x, y)\, dy\, dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} w(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} x w_1(x)\, dx + \int_{-\infty}^{\infty} y w_2(y)\, dy = E(x) + E(y) \end{aligned}$$

5.5.4 Regression and Correlation

If the analyst deals simultaneously with a few random variables, say two, each of them may well be described by their expectations and variances. However, often it is important to know also whether or not one of these variables affects the other in order to take into account the relation between the variables. Quantitative expressions for such a relationship are provided by the regression and correlation.

Let x and y be two random variables. To approach these concepts by way of a specific example, assume that x is the height and y is the diameter of trees in a forest area. We may put in correspondence to every tree a point on the x, y plane, so that the whole population of trees appears as a family of points shown in Fig. 5.7*a*. The expectations of height and diameter are $\bar{x} = E(x)$ and $\bar{y} = E(y)$, and the

variances σ_x^2 and σ_y^2 characterize the dispersion of diameters and heights about their expectations (means).

If we investigate to see whether the diameter depends on the height, we notice that for trees of the same height x the diameter is a random variable. Hence, we can speak only about the dependence of the expected value of diameter, $\bar{y}$, on height, i.e. define the quantity $\bar{y}(x)$ which is the conditional expectation of one variable, given the value of the other, $E(y \mid x)$. This quantity is called the *regression* of that variable on the other.

By measuring $\bar{y}(x)$ at different x we can construct a line that represents the regression of y and x graphically. The dependence of $\bar{x}(y) = E(x \mid y)$, called the regression of x on y, may be obtained in a similar manner.

Thus far we confine ourselves to considering the regression for the most frequent practical situation of linear regression, when the equation of the regression line may be written as

$$\bar{y} = a + b(x - \bar{x}) \tag{5.86}$$

The coefficients a and b are selected such that the concentration of points (x, y) will be highest near the $\bar{y}(x)$ line. This tracing of the

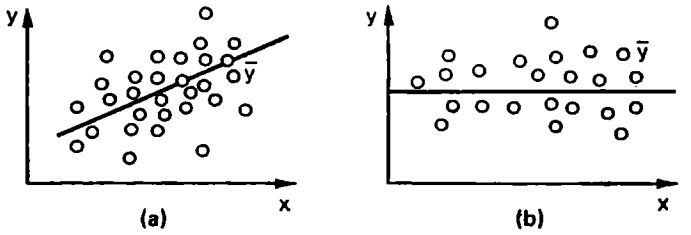


Fig. 5.7. Regression line $y(x)$ for (*a*) dependent and (*b*) independent random variables

line may be obtained with the aid of the method of least squares, as has been done already in Sect. 4.2.

We fix a certain x and denote by $\varepsilon(x) = y - \bar{y}(x) = y - a - b(x - \bar{x})$ the deviation of the ordinate of an experimental point y from the regression line $\bar{y}(x)$. The method of least squares requires that the expectation of the squares of all such deviations

$$J = E[\varepsilon^2(x)] \tag{5.87}$$

should be minimized. We shall achieve this objective by equating the derivatives of J with respect to a and b to zero. This gives us

two equations for the evaluation of these coefficients:

$$\frac{\partial J}{\partial a} = E\left[\frac{\partial \varepsilon^2}{\partial a}\right] = 2E\left[\varepsilon \frac{\partial \varepsilon}{\partial a}\right] = 0$$
$$\frac{\partial J}{\partial b} = E\left[\frac{\partial \varepsilon^2}{\partial b}\right] = 2E\left[\varepsilon \frac{\partial \varepsilon}{\partial b}\right] = 0$$

Observing that $\partial \varepsilon (x)/\partial b = -(x - \bar{x})$ and $\partial \varepsilon (x)/\partial a = -1$, we get

$$E[y - a - b(x - \bar{x})] = 0$$
$$E\{[y - a - b(x - \bar{x})](x - \bar{x})\} = 0$$

From the first equation we find

$$E(y) - a - bE(x - \bar{x}) = 0 \text{ or } a = E(y) = \bar{y} \tag{5.88}$$

The second equation yields

$$E[(y - \bar{y})(x - \bar{x})] - bE[(x - \bar{x})^2] = 0 \tag{5.89}$$

where $E[(x - \bar{x})^2]$ is clearly the variance of x.

We call the quantity

$$\operatorname{cov}(y, x) = E[(y - \bar{y})(x - \bar{x})] \tag{5.90}$$

the *covariance* of x and y. Now the equation for b becomes

$$\operatorname{cov}(y, x) - b\sigma_x^2 = 0$$

whence

$$b = \operatorname{cov}(y, x)/\sigma_x^2 \tag{5.91}$$

Now the regression equation (5.86) may be written as

$$\bar{y}(x) = \bar{y} + \operatorname{cov}(y, x)(x - \bar{x})/\sigma_x^2 \tag{5.92}$$

The covariance of y and x gives a measure of relation of random variables x and y. However, because we have confined ourselves to consideration of the case of linear regression, the covariance will represent a linear relation, i.e. a tendency of one random variable to increase or decrease on average by a linear law as the other variable increases or decreases. For a more vivid representation of the dependence of the covariance on the type of relation between random variables it would be instructive to look at extreme situations.

(1) $\operatorname{cov}(y, x) = 0$. This implies that $b = 0$, $\bar{y}(x) = a$ and the regression line is horizontal (Fig. 5.7*b*). At experimental points, the values of y cluster about the value a independent of x. Consequently, in this case, y and x are independent and said to be uncorrelated.

(2) The variables y and x will be allied most closely if this relation will be linear; all the experimental points will then lie on the regres-

sion line so that $y = \bar{y}(x)$. Substituting y for $\bar{y}$, we rewrite the equation of regression as

$$y - \bar{y} = [\text{cov}(y, x)/\sigma_x^2](x - \bar{x})$$

The variance

$$\sigma_y^2 = E[(y - \bar{y})^2] = [\text{cov}^2(y, x)/\sigma_x^4] E[(x - \bar{x})^2] = \text{cov}^2(y, x)/\sigma_x^2$$

whence

$$\text{cov}(y, x) = \sigma_x \sigma_y \tag{5.93}$$

As can be seen, the covariance varies from 0 to $\sigma_x\sigma_y$ depending on the degree of relation between y and x.

The covariance is somewhat inconvenient to estimate the degree of connection between random variables y and x, as cov (y, x) depends on the variances of these random variables. It would be more convenient to use the *correlation coefficient*

$$r_{yx} = \text{cov}(y, x)/\sigma_x\sigma_y \tag{5.94}$$

that may vary from zero for independent random variables to unity for random variables connected by a linear dependence.

The covariance is helpful in deriving simple relationships for the expected value of a product and the variance of a sum of two random variables. We leave it to the reader to verify that these expressions are as follows:

$$\overline{xy} = \bar{x}\bar{y} + \text{cov}(x, y) \tag{5.95}$$

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2\text{cov}(x, y) \tag{5.96}$$

When random variables x and y are independent,

$$\overline{xy} = \bar{x}\bar{y} \tag{5.97}$$

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) \tag{5.98}$$

5.6 The Multivariate Normal Distribution

5.6.1 Expectation Vector and Dispersion Matrix

If the set of random variables under study consists of more than two variates, say n variates, then the number of numerical parameters characterizing these variates increases in proportion. In addition to the n first moments defining the expectations of the random variates $\bar{x}_i = E(x_i)$ $i = 1, \ldots, n$ we need n^2 central moments representing the variances of each random variable and the covariances for every pair of the random variables. Therefore in multivariate distributions it is convenient to discuss the joint probability

behavior of a set $x_1, \ldots, x_n$ of random variables in terms of the vector random variable (or random vector) written as a column vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = (x_1, \ldots, x_n)^{\mathrm{T}} \tag{5.99}$$

We define the expectation vector as

$$\bar{\mathbf{x}} = E(\mathbf{x}) = \begin{pmatrix} \bar{x}_1 \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_n \end{pmatrix} = (\bar{x}_1, \ldots, \bar{x}_n)^{\mathrm{T}} \tag{5.100}$$

The set of second central moments, i.e. variances

$$\operatorname{var}(x_i, x_i) = E[(x_i - \bar{x}_i)^2], \quad i = 1, \ldots, n \tag{5.101}$$

and covariances

$$\operatorname{cov}(x_i, x_j) = E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)], \quad i, j = 1, \ldots, n; \; i \neq j \tag{5.102}$$

are conveniently written in the form of the *dispersion matrix*

$$\mathbf{D}_{xx} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}}] \tag{5.103}$$

The diagonal elements of this matrix are the variances $\operatorname{var}(x_i) = \sigma_{x_i}^2$ the other elements being the covariances $\operatorname{cov}(x_i, x_j)$. Since $\operatorname{cov}(x_i, x_j) = \operatorname{cov}(x_j, x_i)$, the dispersion matrix is symmetric. Note that some authors call $\mathbf{D}$ the *variance*, *covariance*, or *variance-covariance matrix*.

Example 5.12. Let $\mathbf{x} = (u, v)^{\mathrm{T}}$ be a two-dimensional random vector. Then the expectation vector is

$$\bar{\mathbf{x}} = E(\mathbf{x}) = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}, \text{ where } \bar{u} = E(u) \text{ and } \bar{v} = E(v)$$

The dispersion matrix is

$$\mathbf{D}_{xx} = E\left[\begin{pmatrix} u - \bar{u} \\ v - \bar{v} \end{pmatrix}(u - \bar{u}, \; v - \bar{v})\right]$$

$$= E\begin{bmatrix} (u-\bar{u})^2 & (u-\bar{u})(v-\bar{v}) \\ (v-\bar{v})(u-\bar{u}) & (v-\bar{v})^2 \end{bmatrix} = \begin{pmatrix} \operatorname{var}(u) & \operatorname{cov}(u, v) \\ \operatorname{cov}(v, u) & \operatorname{var}(v) \end{pmatrix}$$

The covariances $\operatorname{cov}(x_i, y_i)$ of the members of two sets of random variables $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, one of one

set with one of another, may be manipulated in terms of the *cross-covariance matrix*

$$\mathbf{C}(\mathbf{x}, \mathbf{y}) = E(\mathbf{x}, \mathbf{y})^{\mathrm{T}} - E(\mathbf{x})\, E(\mathbf{y}) \tag{5.104}$$

It is important to note that

$$\mathbf{C}(\mathbf{x}, \mathbf{y}) = \mathbf{C}(\mathbf{y}, \mathbf{x})^{\mathrm{T}} \tag{5.105}$$

5.6.2 Normal Distribution of a Random Vector

In Sect. 5.3 we noted that the normal distribution law stands separately among other distribution laws. This special role of the normal law holds for multivariate vectors. This role is underlaid by several reasons as follows.

1. Many physical phenomena obey this law.
2. This law is the limiting one for many other laws.
3. This distribution has a comparatively simple mathematical description, for it is completely described by solely the first and second moments, i.e. by the expectation vector and the dispersion matrix. Other distributions require moments of higher order in their description.
4. Linear transformations of random vectors distributed by a normal law result in new random vectors that also obey a normal distribution. In other words, if $\mathbf{x}$ and $\mathbf{y}$ are normally distributed random vectors, and $\mathbf{A}$ and $\mathbf{B}$ are matrices, then the random vector $\mathbf{z} = \mathbf{Ax} + \mathbf{By}$ will be distributed by a normal law too with readily computable expectation and the dispersion matrix.

Example 5.13. Let $\mathbf{x}$ be a random n-vector, $\mathbf{A}$ an $r \times n$ matrix, and $\mathbf{z} = \mathbf{Ax}$ a random r-vector. The expectation of vector $\mathbf{z}$ is

$$\bar{\mathbf{z}} = E(\mathbf{Ax}) = \mathbf{A}E(\mathbf{x}) = \mathbf{A}\bar{\mathbf{x}}$$

and its dispersion matrix

$$\mathbf{D}(\mathbf{z}, \mathbf{z}) = E[(\mathbf{z} - \bar{\mathbf{z}})(\mathbf{z} - \bar{\mathbf{z}})^{\mathrm{T}}] = E[\mathbf{A}(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}]$$
$$= \mathbf{A}\mathbf{D}_{xx}\mathbf{A}^{\mathrm{T}}$$

We say that a random n-vector $\mathbf{x}$ is distributed to a Gaussian (normal) law, if its probability density function $w(\mathbf{x}) = w(x_1, \ldots, x_n)$ is given by the expression

$$w(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \mid \mathbf{D}_{xx} \mid}} \exp\left[-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{D}_{xx}^{-1} (\mathbf{x} - \bar{\mathbf{x}})\right] \tag{5.106}$$

where $\mid \mathbf{D}_{xx} \mid = \det \mathbf{D}_{xx}$ is the determinant of the dispersion matrix $\mathbf{D}_{xx}$.

If x is a scalar, then $n = 1$, $(x - \bar{x})^{\mathrm{T}} = (x - \bar{x})$, $\mathbf{D}_{xx} = \sigma^2$, $\mid \mathbf{D}_{xx} \mid = \sigma^2$, $\mathbf{D}_{xx}^{-1} = 1/\sigma^2$, and (5.106) reduces to (5.40) for the normal distribution of a unidimensional random variate.

5.6.3 Normal Distribution of Two Random Vectors

Let $\mathbf{u} = (u_1 \ \ldots \ u_r)^{\mathrm{T}}$ and $\mathbf{v} = (v_1 \ \ldots \ v_k)^{\mathrm{T}}$ be two random r- and k-vectors with $r + k = n$. Denote by $\mathbf{x} = (\mathbf{u}, \mathbf{v}) = (u_1, \ldots, u_r, v_1, \ldots, v_k)^T$ the random n vector with components $\mathbf{u}$ and $\mathbf{v}$. For this case (5.106) yields the joint distribution $w(\mathbf{x}) = w(\mathbf{u}, \mathbf{v})$ of $\mathbf{v}$ and $\mathbf{u}$. The expectation and dispersion matrix in (5.106) will be

$$\bar{\mathbf{x}} = E(\mathbf{x}) = E\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} = \begin{pmatrix}\bar{\mathbf{u}}\\ \bar{\mathbf{v}}\end{pmatrix}, \quad \bar{\mathbf{u}} = E(\mathbf{u}), \quad \bar{\mathbf{v}} = E(\mathbf{v}) \tag{5.107}$$

$$\mathbf{D}_{xx} = E\left\{\begin{pmatrix}\mathbf{u}-\bar{\mathbf{u}}\\ \mathbf{v}-\bar{\mathbf{v}}\end{pmatrix}[(\mathbf{u}-\bar{\mathbf{u}})^{\mathrm{T}}, \quad (\mathbf{v}-\bar{\mathbf{v}})]^{\mathrm{T}}\right\} = \begin{pmatrix}\mathbf{D}_{uu} & \mathbf{C}_{uv}\\ \mathbf{C}_{vu} & \mathbf{D}_{vv}\end{pmatrix} \tag{5.108}$$

where $\mathbf{D}_{uu}$ and $\mathbf{D}_{vv}$ are the dispersion matrices of $\mathbf{u}$ and $\mathbf{v}$, and $\mathbf{C}_{uv}$ and $\mathbf{C}_{vu} = \mathbf{C}_{uv}^{\mathrm{T}}$ are the cross-covariance matrices.

If we represent the inverse matrix $\mathbf{D}_{xx}^{-1}$ in the form

$$\mathbf{D}_{xx}^{-1} = \begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^{\mathrm{T}} & \mathbf{C}\end{pmatrix} \tag{5.109}$$

where $\mathbf{A}$ and $\mathbf{C}$ are symmetric matrices of type $r \times r$ and $k \times k$ respectively, and $\mathbf{B}$ is an $r \times k$ matrix, then the component matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ may be determined from the condition $\mathbf{D}_{xx}\mathbf{D}_{xx}^{-1} = \mathbf{I}$ as

$$\begin{aligned}\mathbf{A} &= (\mathbf{D}_{uu} - \mathbf{C}_{uv}\mathbf{D}_{vv}^{-1}\mathbf{C}_{vu})^{-1}\\ \mathbf{C} &= (\mathbf{D}_{vv} - \mathbf{C}_{vu}\mathbf{D}_{uu}^{-1}\mathbf{C}_{uv})^{-1}\\ \mathbf{B} &= \mathbf{A}\mathbf{C}_{uv}\mathbf{D}_{vv}^{-1} = -\mathbf{D}_{uu}^{-1}\mathbf{C}_{uv}\mathbf{C}\end{aligned} \tag{5.110}$$

In this nomenclature the joint distribution of two random vectors $\mathbf{u}$ and $\mathbf{v}$ assumes the form

$$w(\mathbf{u}, \mathbf{v}) = [(2\pi)^{r+k} \mid \mathbf{D}_{xx} \mid]^{-1/2} \times \exp\left[-\frac{1}{2}\begin{pmatrix}\mathbf{u}-\bar{\mathbf{u}}\\ \mathbf{v}-\bar{\mathbf{v}}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^{\mathrm{T}} & \mathbf{C}\end{pmatrix}\begin{pmatrix}\mathbf{u}-\bar{\mathbf{u}}\\ \mathbf{v}-\bar{\mathbf{v}}\end{pmatrix}\right] \tag{5.111}$$

5.6.4 Conditional Normal Distribution

We shall consider the conditional p.d.f. for two vectors $\mathbf{u}$ and $\mathbf{v}$ by analogy with bivariate distributions outlined in Sect. 5.5. In view of (5.82) the conditional p.d.f. is defined as

$$w(\mathbf{u} \mid \mathbf{v}) = w(\mathbf{u}, \mathbf{v})/w(\mathbf{v}) \tag{5.112}$$

If we recall that

$$w(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^k \mid \mathbf{D}_{vv} \mid}} \exp\left[-\frac{1}{2}(\mathbf{v}-\bar{\mathbf{v}})^{\mathrm{T}}\mathbf{D}_{vv}^{-1}(\mathbf{v}-\bar{\mathbf{v}})\right]$$

and (5.111), we get

$$w(\mathbf{u} \mid \mathbf{v}) = |(2\pi)^r \mid \mathbf{D}_{xx} \mid / \mid \mathbf{D}_{vv} \mid|^{-1/2} \times \exp\left[-\frac{1}{2} \begin{pmatrix} \mathbf{u}-\bar{\mathbf{u}} \\ \mathbf{v}-\bar{\mathbf{v}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{C}-\mathbf{D}_{vv}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{u}-\bar{\mathbf{u}} \\ \mathbf{v}-\bar{\mathbf{v}} \end{pmatrix} \right]$$

After some matrix manipulations these expressions can be written in a more succinct form

$$w(\mathbf{u} \mid \mathbf{v}) = |(2\pi)^r \mid \mathbf{Q} \mid|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{u}-\mathbf{m})^{\mathrm{T}} \mathbf{Q}^{-1} (\mathbf{u}-\mathbf{m}) \right] \quad (5.113)$$

where

$$\mathbf{m} = E(\mathbf{u} \mid \mathbf{v}) = \bar{\mathbf{u}} + \mathbf{C}_{uv}\mathbf{D}_{vv}^{-1}(\mathbf{v} - \bar{\mathbf{v}}) \quad (5.114)$$

is the conditional expectation, and

$$\mathbf{Q} = \mathbf{D}_{u|v} = \mathbf{D}_{uu} - \mathbf{C}_{uv}\mathbf{D}_{vv}^{-1}\mathbf{C}_{vu} \quad (5.115)$$

is the conditional dispersion matrix.

5.7 Stochastic Processes in Discrete Time

5.7.1 Introduction

Suppose that a system subjected to some external conditions changes its state at random at time instants that will be labeled conditionally by 0, 1, etc. The time intervals between these instants will be referred to as *steps*. After step n the system assumes the state x_n that may be treated as an outcome of an experiment, i.e. as an element of a feasible set of outcomes for this step, $Z_n = \{z_0^n, z_1^n, \ldots, z_L^n\}$.

We assume that the set Z_0 consists of one element z_0^0 representing the initial state of the system. The sequence of states x_0, x_1 etc. which the system assumes is called a *random process with discrete time*. This random process will be specified if at each step we know the probability distribution for the outcomes, i.e. states, which the system may assume at the next step. In general this probability distribution depends on the states the system occupied in all previous steps, and for any n is defined by the set of conditional probabilities

$$p(x_{n+1} = z_k^{n+1} \mid x_0, \ldots, x_n), \quad k = 0, 1, \ldots, L \quad (5.116)$$

where

$$\sum_{k=0}^{L} p(x_{n+1} = z_k^{n+1} \mid x_0, \ldots, x_n) = 1 \quad (5.117)$$

This random process can continue indefinitely long if it will not be interrupted at some step. In practice, we always interrupt the process after a certain number of steps, i.e. we normally deal with

finite or *multiple step processes.* However, infinite processes are useful approximations of actual finite processes, because they are often simpler from the analytical standpoint.

A random process whose behavior cannot be affected and, consequently, whose probability distribution (5.116) cannot be influenced will be referred to as *uncontrolled process.* Depending on the type of the distribution (5.116), random processes are categorized into several types, three of which are encountered more often than others — processes with independent values, processes with independent trials, and Markov chains.

A finite random process is said to have independent values if the probability distribution function in the realization sample space at each step is independent of the results at previous steps,

$$p\,(x_n = z_k^n \mid x_0, \ldots, x_{n-1}) = p\,(x_n = z_k^n) \qquad (5.118)$$

This probability distribution is conveniently denoted simply by $p_n\,(z)$, where $z \in Z_n$, $n = 0, 1$, etc.

An important property of a process with independent values is that any sequence of states $x_0, x_1, \ldots, x_n$ may be viewed as a collection of independent outcomes. Consequently, the probability of this sequence will be equal to the product of the probabilities of outcomes constituting this sequence, which may be written as

$$p\,(x_0, x_1, \ldots, x_n) = p_0\,(x_0)\,p_1\,(x_1) \ldots p_n\,(x_n) \qquad (5.119)$$
$$x_i \in Z_i,\ i = 0, 1, \ldots, n$$

An important particular case of processes with independent values is processes of independent trials. A *process of independent trials* is the one with independent values having the sample space $Z = \{z_0, z_1, \ldots, z_L\}$ and the probability distribution in this space is the same at each step, i.e.

$$p_n\,(z) = p_m\,(z) = p\,(z) \qquad (5.120)$$

for every $z \in Z$ and any m and n. In other words, the process of independent trials is a multiple repetition of one experiment under the same conditions.

Like in the case of processes with independent variables, the probability of a particular sequence of outcomes $x_0, x_1, \ldots, x_n$ in a process of independent trials equals the product of the probabilities of individual outcomes.

A random process is said to be a *Markov chain* if its sample space $Z = \{z_0, z_1, \ldots, z_L\}$ is the same at each step, and the distribution of probabilities at each step depends only on the outcome at the preceding step so that

$$p\,(x_{n+1} = z \mid x_0, \ldots, x_n) = p\,(x_{n+1} = z \mid x_n) \qquad (5.121)$$

Suppose that at step n the system assumes the state $z_i \in Z$. Then the probability distribution (5.121) defines the probability of the system occupying state $z_j \in Z$ at step $n + 1$. In the general case, this probability depends on the number of the step and may be denoted by $p_{ij}(n)$. Thus the probabilities

$$p_{ij}(n) = p(x_{n+1} = z_j \mid x_n = z_i) \tag{5.122}$$

satisfying the conditions

$$p_{ij}(n) \geqslant 0, \quad \sum_{j=0}^{L} p_{ij}(n) = 1 \tag{5.123}$$

are the probabilities of transition at step n from state z_i to state z_j.

When the transition probabilities $p_{ij}(n)$ depend on the step number n, the Markov chain is said to be non-stationary. In applications, however, the predominant Markov chains are those with transition probabilities identical at any step.

5.7.2 The Two-State Process. The Binomial Distribution

Consider in more detail the process of independent trials each of which may or may not result in some event S. One outcome of this experiment we call a success and denote by 1, the other a failure and denote by 0.

Denote by p the probability of success in one experiment, and by $q = 1 - p$ the probability of failure. Suppose now we wish to know the probability $w(r, n, p)$ of obtaining r successes in n trials. In the above nomenclature, obtaining r successes in n trials will correspond to a sequence of n outcomes consisting of r units and $n - r$ zeros. Because the results of individual trials are independent, the probability of this sequence equal to the product of probabilities of individual trials will be given by

$$p^r q^{n-r} \tag{5.124}$$

The total number of different sequences of n elements containing r units equals the number of combinations of n elements taken r at a time, i.e.

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!} \tag{5.125}$$

All these sequences are inconsistent (i.e. if one of them occurs, it prevents the occurrence of another), therefore the probability that one of them occurs is

$$w(r, n, p) = \binom{n}{r} p^r q^{n-r} \tag{5.126}$$

It is an easy matter to verify that under the circumstances the total probability of occurrence of any number of events from 0 to n equals unity.

The distribution defined by (5.126) is the *binomial distribution.* For $n = 10$ and $p = 1/5$ the plot of this distribution versus r parameters is shown in Fig. 5.8.

It is often required to determine the value of r at which the bino-

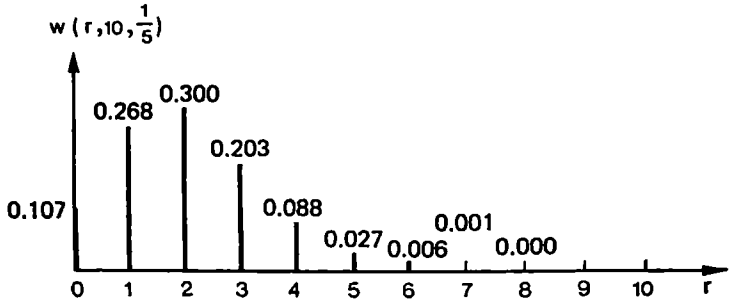


Fig. 5.8. Binomial distribution

mial distribution is a maximum. To solve this problem we first determine the ratio of two adjacent ordinates in the binomial distribution

$$\frac{w(r,\ n,\ p)}{w(r-1,\ n,\ p)} = \binom{n}{r} p^r q^{n-r} \Big/ \binom{n}{r-1} p^{r-1} q^{n-r+1} = (n-r+1)\, p/rq \tag{5.127}$$

This ratio will not be less than unity if $r \leqslant (n+1)\, p$. Consequently the function $w\ (r,\ n,\ p)$ is a maximum at

$$r = \text{ent}\ (n+1)\, p \tag{5.128}$$

where ent x denotes the integer part of x.

The numerical characteristics of a binomial distribution are the expectation and variance, respectively,

$$\mu_r = np, \quad \sigma_r^2 = npq \tag{5.129}$$

5.7.3 The Poisson Distribution

The values of $w\ (r,\ n,\ p)$ are hard to calculate at large n. One of approximate methods that facilitate these calculations consists in using the Poisson distribution in place of the binomial distribution.

Let $w\ (r,\ n,\ p)$ be a binomial distribution. Assuming $np = a = \text{const}$, it is an easy matter to verify that the limit of this distribution for n tending to infinity and p tending to zero is

$$\lim_{n\to\infty,\ p\to 0} w\ (r,\ n,\ a/n) = a^r e^{-a}/r! \tag{5.130}$$

We shall denote this distribution by $w(r, a)$ and call it the *Poisson distribution*. Thus,

$$w(r, a) = a^r e^{-a}/r! \tag{5.131}$$

A characteristic feature of this distribution is that its expectation and variance are both equal to a.

The Poisson distribution is a good approximation for the binomial distribution at small p. For example, at $p \leqslant 0.01$ the binomial distribution may be replaced by the Poisson distribution beginning with $n = 10$.

Although we arrived at the Poisson distribution as the limiting case of the binomial distribution, it belongs to a class of stochastic processes categorized as *discrete-state continuous time processes*. Consider a sequence of events that follow one another over random intervals of time. The number of such events occurring in the interval τ will be a random variable with expectation $\lambda\tau$, where λ is the mean number of events per unit time called the *intensity* or *rate* of the process.

Let us determine the probability that exactly r events will occur in the interval τ. We assume that these events are not frequent, so that the interval τ may be divided into small subintervals $\Delta\tau$ in each of which not more than one event may occur. The total number of such subintervals is $n = \tau/\Delta\tau$, and the probability of an event occurring in one of these subintervals $\Delta\tau$ will be $p = \lambda\tau/n$, so that $np = \lambda\tau$. We assume that an event occurs in an interval $\Delta\tau$ independently of whether or not it occurred in another interval, therefore we have a process of independent trials with probability density $w(r, n, p)$, where $np = \lambda\tau$.

Our assumption that an interval $\Delta\tau$ may house a single event is justified only if the length of this interval is very small, i.e. in the limit as $\Delta\tau \to 0$. This is accompanied by $n \to \infty$ and $p \to 0$ with $np = \lambda\tau = \text{const}$, therefore the discrete-state continuous-time process obeys the Poisson distribution with probability density function

$$w(r, \lambda\tau) = (\lambda\tau)^r e^{-\lambda\tau}/r! \tag{5.132}$$

5.7.4 The Exponential Distribution. The Concept of Reliability

Formulations of many problems are centered on the probability distribution of time intervals elapsing between random events. From (5.132) at $r = 0$ and $r = 1$ it follows that $e^{-\lambda\tau}$ is the probability that no event occurs in the interval of length τ, and $\lambda\tau e^{-\lambda\tau}$ is the probability that only one event occurs in the interval τ.

The probability that more than one event occurs in the interval of length τ is

$$1 - (e^{-\lambda\tau} + \lambda\tau e^{-\lambda\tau}) = 1 - \{[1 - \lambda\tau + (\lambda\tau)^2/2! - \ldots] - \lambda\tau\,]1 - \lambda\tau + (\lambda\tau)^2/2! - \ldots]\} = (\lambda\tau)^2/2 + \ldots \quad (5.133)$$

This expression suggests that the probability is proportional to $(\lambda\tau)^2$ which implies that at small $\lambda\tau$ we may neglect the probability that two or more events occur in the interval τ. Then, the probability of obtaining one event in the interval τ is

$$F(\lambda, \tau) = 1 - e^{-\lambda\tau}, \quad \tau \geqslant 0 \quad (5.134)$$

This expression essentially defines the probability that the time within which one event may occur will not exceed τ. The density of this probability, i.e. the probability that this event will occur at time τ, can be obtained by differentiating (5.134) with respect to τ, viz.,

$$w(\lambda, \tau) = \lambda e^{-\lambda\tau}, \quad \tau \geqslant 0 \quad (5.135)$$

Now that we know the probability density function for τ, it will be easy to determine the average time interval during which no event will occur, namely,

$$\tau_{av} = \int_0^\infty \tau w(\lambda\tau)\, d\tau = 1/\lambda \quad (5.136)$$

The formulas (5.135) and (5.136) define the exponential distribution which plays an important role in reliability theory [24].

Here is a brief excursion into reliability theory. Suppose that a system under test consists of a large number of subsystems or components each of which may fail after time t. In most cases of engineering interest components fail independently of one another and a sequence of failures may be viewed as a process of independent events with *failure rate* λ and, according to (5.136), *mean time to failure* (MTTF) $1/\lambda$.

In reliability engineering, the reliability of equipment is usually characterized by the quantity $h(\tau)$ called the *hazard rate* and defined as the probability density of failure at time τ provided there were no failures until this instant of time.

Let us determine the quantity $h(\tau)\,\Delta\tau$ that gives the probability that a failure occurs in the interval $\Delta\tau$ provided that there were no failures during the time interval τ. By the formula for conditional probability we have

$$h(\tau)\,\Delta\tau = w(\lambda, \tau)\,\Delta\tau/[1 - F(\lambda, \tau)] \quad (5.137)$$

whence

$$h(\tau) = w(\lambda, \tau)/[1 - F(\lambda, \tau)] \quad (5.138)$$

For the exponential distribution we have

$$h(\tau) = \lambda e^{-\lambda\tau}/e^{-\lambda\tau} = \lambda \tag{5.139}$$

that is, the hazard rate is a constant equal to the failure rate.

Experience with the failure behavior of electronic components indicates that a typical hazard rate curve has the *bathtub* form shown in Fig. 5.9. The initial portion of this curve exhibits a decreasing hazard rate normally attributed to design or manufacturing defects. This portion is sometimes called the "burn-in period", though there are several other names for this period such as debugging period, infant mortality period, or break-in period.

Fig. 5.9. The bathtub curve

The constant part of this bathtub hazard rate curve is called the "useful life period". It begins just after the burn-in period and ends just before the "wear-out period". The constant failure rate means that this period is described by an exponential distribution. Failures during this period occur randomly.

The wear-out period begins when the system or component has aged or bypassed its useful operating life. Therefore the number of failures during this period begins to increase.

The high failure rate at the initial period can be reduced by aging the component prior to operation. On the other hand, the requirements on the component performance become ever more severe, so the component may become obsolete in its useful life period and be replaced by an updated version. Therefore, in many situations the hazard rate may be deemed constant and equal over its entire operating life and invoke the exponential law in reliability estimation.

5.7.5 Markov Chains

Let $Z = \{z_1, \ldots, z_L\}$ be the sample space of an experiment or state space of some system, identical for every step of the stochastic process under study. Then the states of the system are conveniently indicated by the labels of z_j understanding j as the state $z_j \in Z$. Accordingly, the state space will be represented by the index set $J = \{1, \ldots, L\}$.

Denote by $\pi_n = (\pi_n^{(1)}, \ldots, \pi_n^{(L)})$ the probability distribution on the set J for step n; $\pi_n^{(j)}$ gives the probability that the system is at state j at step n. As has been pointed out a stochastic process is a homogeneous Markov chain if its transition probabilities p_{ij} depend only on state i at the preceding step and are identical at every step.

The transition probabilities can be exhibited in two different ways. The first way is that of a square array, called the transition (probability) matrix. For a Markov chain with states z_1, z_2, and z_3, this array is written as

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \tag{5.140}$$

The condition (5.123) which the transition probabilities must obey implies that the elements in a row add up to 1, corresponding to the fact that if we start at a given state we must reach some state after three steps.

The second way to show the transition probabilities is by a *transition diagram*. Such a diagram is illustrated for a system with three states in Fig. 5.10*a*. This diagram is a graph whose vertices represent

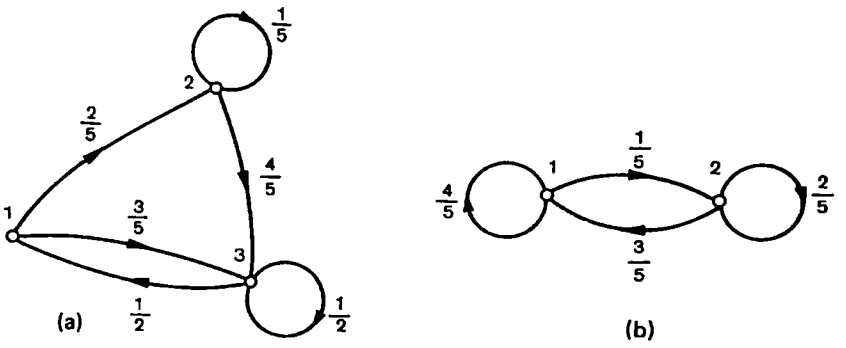


Fig. 5.10. Transition graphs for a Markov chain

states of the system and directed edges, or arcs, represent feasible transitions from state to state. The associated probabilities are indicated by numbers shown next to each arc. In agreement with the condition (5.123) the sum of the probabilities for arcs emanating from any vertex of the graph must be unity.

Many processes in various fields of science and engineering may be modeled by Markov chain processes. In social science, Markov chains are helpful in studying variations of the social or professional structure of population, migration of people, and the like. In biology, Markov processes model the evolution of certain species. In physics, Markov chain processes are helpful in studying diffusion of gases. Engineering makes good use of Markov models for certain processes of data transmission, flow lines, and reliability engineering.

The kind of problem in which we are most interested in the study of Markov chains is how the probability distribution of states π_n will change when the system moves one step. Denote by T the event

of the system occupying state j at step $n+1$. Denote the probability of this event by $P(T) = \pi_{n+1}^{(j)}$. Let S_i be the event of the system being in state i at step n, so that $P(S_i) = \pi_n^{(i)}$. The transition probability p_{ij} implies the probability that the system arrives at state j after step $n+1$ provided that it occupied state i at step n, i.e. $p_{ij} = p(T \mid S_i)$. From (5.25) it follows that

$$\pi_{n+1}^{(j)} = \sum_{i=1}^{L} \pi_n^{(i)} p_{ij}, \quad j = 1, \ldots, L \tag{5.141}$$

If we know the initial distribution of probabilities, then using this formula we can step by step determine variations in the probability distribution of the system.

Example 5.14. A factory is manufacturing TV sets of a certain type. At the end of each fiscal year the factory is faced with the following alternative: TV sets sell well, state 1, or TV sets do not sell, state 2. The situation on the market changes in time so that if the factory is in state 1 the probability to remain in this state to the end of the year is 4/5. If the factory is in state 2, then usually some measures are taken (quality improvement, additional advertisement, etc.) to improve the selling situation, so that the probability to appear in state 1 by the end of the year is 3/5. The situation may be represented by a Markov chain with transition matrix

$$\mathbf{P} = \begin{pmatrix} 4/5 & 1/5 \\ 3/5 & 2/5 \end{pmatrix}$$

The associated transition diagram is shown in Fig. 5.10*b*.

Example 5.15. Assume that the factory of the previous example is initially in state 1, i.e. the initial probability distribution is $\pi_0 = (1, 0)$. Making use of (5.141) and the transition matrix, we find for the first step

$$\pi_1^{(1)} = \pi_0^{(1)} p_{11} + \pi_0^{(2)} p_{21} = 1 \times 4/5 + 0 \times 3/5 = 0.8$$
$$\pi_1^{(2)} = \pi_0^{(1)} p_{12} + \pi_0^{(2)} p_{22} = 1 \times 1/5 + 0 \times 2/5 = 0.2$$

and for the second step

$$\pi_2^{(1)} = 0.76 \quad \text{and} \quad \pi_2^{(2)} = 0.24$$

If we continue these calculations, we obtain successful variation of the probability distribution (Table 5.1) for the initial state (1, 0). The data summarized in the table indicate that for greater n the value of $\pi_n^{(1)}$ approaches 0.75 and that of $\pi_n^{(2)}$ approaches 0.25.

The table also shows the probability distribution at successful steps for the initial state $\pi_0 = (0, 1)$. In this case, with the growth of n the values of $\pi_n^{(1)}$ and $\pi_n^{(2)}$ approach the same figures, 0.75 and

Table 5.1

Initial state n	$\pi_0 = (1, 0)$		$\pi_0 = (0, 1)$	
	$\pi_n^{(1)}$	$\pi_n^{(2)}$	$\pi_n^{(1)}$	$\pi_n^{(2)}$
0	1	0	0	1
1	0.8	0.2	0.6	0.4
2	0.76	0.24	0.72	0.28
3	0.752	0.248	0.744	0.256
4	0.7504	0.2496	0.7488	0.2512
5	0.75008	0.24992	0.74976	0.25024
⋮	⋮	⋮	⋮	⋮

0.25, found for the initial state $\pi_0 = (1, 0)$. Therefore, in this Markov chain the transition probabilities become independent of the initial state of the system after a large number of steps.

If for $n \to \infty$ the probability distribution of a Markov chain converges to a limiting distribution, then this distribution defines the equilibrium regime of this chain process, which does not depend in any way on the initial conditions. The system with equilibrium distribution is said to be *statically stable*, or *stationary* and the associated Markov process is called *ergodic*.

Denote by $\pi = (\pi^{(1)}, \ldots, \pi^{L})$ the equilibrium distribution of an ergodic Markov chain. The components $\pi^{(j)}$ of this distribution may be derived by Eqs. (5.141) which for $n \to \infty$ take the form

$$\pi^{(j)} = \sum_{i=1}^{L} \pi^{(i)} p_{ij}, \quad j = 1, \ldots, L \tag{5.142}$$

However, not all L equations in (5.142) are linearly independent, for the probabilities $\pi^{(j)}$ are related by the normalizing condition

$$\sum_{j=1}^{L} \pi^{(j)} = 1 \tag{5.143}$$

Therefore, to determine the L unknown components $\pi^{(j)}$ of the probability distribution π it suffices to take any $L - 1$ equations of (5.142) and solve them simultaneously with equation (5.143).

Example 5.16. For the case of transition matrix given in Example 5.14, the first of equations (5.142) has the form

$$\pi^{(1)} = \pi^{(1)} p_{11} + \pi^{(2)} p_{21} = 4\pi^{(1)}/5 + 3\pi^{(2)}/5$$

Equation (5.143) yields $\pi^{(1)} + \pi^{(2)} = 1$. Solving these equations together, we get $\pi^{(1)} = 0.75$ and $\pi^{(2)} = 0.25$.

The sequential transition of an ergodic Markov process from the initial state to the equilibrium distribution is described by a sequence of probability distributions π_n for $n = 1, 2$, etc. For a given initial distribution π_0, this description may be obtained by sequential application of Eq. (5.141) as illustrated in Example 5.15. However, this objective may be attained in some other way.

Consider the probabilities of transition to state j from state i in l steps

$$p_{ij}^{(l)} = p\,(x_{n+l} = j \mid x_n = i), \quad n = 0, 1, \ldots \tag{5.144}$$

These probabilities yield the probability distributions π_l from π_0 by a formula similar to (5.141), namely,

$$\pi_l^{(j)} = \sum_{i=1}^{L} \pi_0^{(i)} p_{ij}^{(l)}, \quad j = 1, \ldots, L \tag{5.145}$$

In order to establish the relation of the transition probabilities $p_{ij}^{(l)}$ with the one-step probabilities p_{ij} we let $l = l_1 + l_2$ to consider the transition of the system from π_0 to π_l in two steps, first to π_{l_1} then to $\pi_l = \pi_{l_1+l_2}$ so that

$$\pi_l^{(j)} = \sum_{k=1}^{L} \pi_{l_1}^{(k)} p_{kj}^{(l_2)} \tag{5.146}$$

where

$$\pi_{l_1}^{(k)} = \sum_{i=1}^{L} \pi_0^{(i)} p_{ik}^{(l_1)} \tag{5.147}$$

Substituting $\pi_{l_1}^{k}$ into (5.146) gives

$$\pi_l^{(j)} = \sum_{k=1}^{L} p_{kj}^{(j_2)} \sum_{i=1}^{L} \pi_0^{(i)} p_{ik}^{(l_1)} = \sum_{i=1}^{L} \pi_0^{(i)} \sum_{k=1}^{L} p_{ik}^{(l_1)} p_{kj}^{(l_2)} \tag{5.148}$$

Comparing (5.148) and (5.145) yields

$$p_{ij}^{(l_1+l_2)} = \sum_{k=1}^{L} p_{ik}^{(l_1)} p_{kj}^{(l_2)} \tag{5.149}$$

For subsequent computation of transition probabilities $p_{ij}^{(1)}$, $p_{ij}^{(2)}$, . . . it is convenient to let $l_1 = l - 1$ and $l_2 = 1$ in (5.149). Then

$$p_{ij}^{(l)} = \sum_{h=1}^{L} p_{ih}^{(l-1)} p_{hj} \tag{5.150}$$

Letting in succession $l = 1, 2, \ldots$ and observing that by definition

$$p_{ik}^{0} = 1, \quad i = k \qquad (5.151)$$
$$= 0, \quad i \neq k$$

we have

$$p_{ij}^{(1)} = p_{ij}$$
$$p_{ij}^{(2)} = \sum_{k=1}^{L} p_{ik} p_{kj} \qquad (5.152)$$
$$\ldots\ldots\ldots$$

Equation (5.150) may be generalized over the case of a Markov chain with transition probabilities dependent on the number of step. Under the circumstances $p_{ij}(n, n_1)$ for $n < n_1$ is the probability that at step n_1 the system appears in state j if it was in state i at step n, namely,

$$p_{ij}(n, n_1) = p(x_{n_1} = j \mid x_n = i), \quad n < n_1 \qquad (5.153)$$

By analogy with (5.145) we have

$$\pi_{n_1}^{(j)} = \sum_{i=1}^{L} \pi_n^{(j)} p_{ij}(n, n_1), \quad n < n_1 \qquad (5.154)$$

On the other hand, considering some n' satisfying the condition $n < n' < n_1$, we may represent $\pi_{n_1}^{(j)}$ in the form

$$\pi_{n_1}^{(j)} = \sum_{k=1}^{L} \pi_{n'}^{(k)} p_{kj}(n', n_1) = \sum_{k=1}^{L} p_{kj}(n', n_1) \sum_{i=1}^{L} \pi_n^{(i)} p_{ik}(n, n')$$
$$= \sum_{i=1}^{L} \pi_n^{(i)} \sum_{k=1}^{L} p_{ik}(n, n') p_{kj}(n', n_1) \qquad (5.155)$$

Comparison of (5.154) and (5.155) leads to

$$p_{ij}(n, n_1) = \sum_{k=1}^{L} p_{ik}(n, n') p_{kj}(n', n_1) \qquad (5.156)$$

for any $n < n' < n_1$. This expression is known as the *Chapman-Kolmogorov equation.*

5.7.6 Gaussian Markov Processes

Markov processes play a key role among various types of stochastic processes by about the same reasons that single out the normal distribution of all other distribution laws. These reasons are high practical recurrence and a comparatively simple mathematical des-

cription. The above Markov chains are a particular case of Markov processes. The relevant definitions will be given below.

Let $\mathbf{x}(t)$ be a scalar or vector stochastic process, i.e. a function of time whose values are random vectors for any time t. Consider a sequence of time instants $t_0 < t_1 < \ldots < t_m < t_{m+1}$. Denote by $\mathbf{x}_i$ the value of $\mathbf{x}(t)$ at time t_i, $i = 0, \ldots, m+1$. The process $\mathbf{x}(t)$ is Markovian if the conditional probability density for $\mathbf{x}(t)$ satisfies the condition

$$w(\mathbf{x}_{m+1} \mid \mathbf{x}_m, \ldots, \mathbf{x}_0) = w(\mathbf{x}_{m+1} \mid \mathbf{x}_m) \tag{5.157}$$

If the time instant t_m represents the "present" and time instants $t_{m-1}, \ldots, t_0$ the "past", then the probability law describing the behavior of the process in the future, i.e. at t_{m+1} depends only upon the current state $\mathbf{x}_m$ of the process and does not depend upon its behavior in the past. This memoryless behavior is referred to as the *Markovian property*.

Another method of describing a Markov process consists in defining the joint probability density function $w(\mathbf{x}_m, \ldots, \mathbf{x}_0)$. By virtue of (5.81) we represent the joint probability density as

$$\begin{aligned} w(\mathbf{x}_m, \ldots, \mathbf{x}_0) &= w(\mathbf{x}_m \mid \mathbf{x}_{m-1}, \ldots, \mathbf{x}_0)\, w(\mathbf{x}_{m-1}, \ldots, \mathbf{x}_0) \\ &= w(\mathbf{x}_m \mid \mathbf{x}_{m-1})\, w(\mathbf{x}_{m-1}, \ldots, \mathbf{x}_0) \end{aligned}$$

Transforming $w(\mathbf{x}_{m-1}, \ldots, \mathbf{x}_0)$ in a similar way, we get

$$w(\mathbf{x}_m, \ldots, \mathbf{x}_0) = w(\mathbf{x}_m \mid \mathbf{x}_{m-1})\, w(\mathbf{x}_{m-1} \mid \mathbf{x}_{m-2})\, w(\mathbf{x}_{m-2}, \ldots, \mathbf{x}_0)$$

Continuing this process leads us to

$$\begin{aligned} w(\mathbf{x}_m, \ldots, \mathbf{x}_0) = w(\mathbf{x}_m \mid \mathbf{x}_{m-1})\, w(\mathbf{x}_{m-1} \mid \mathbf{x}_{m-2}) \ldots w(\mathbf{x}_1 \mid \mathbf{x}_0) \\ \times w(\mathbf{x}_0) \end{aligned} \tag{5.158}$$

If the conditional probability density $w[\mathbf{x}(t) \mid \mathbf{x}(\tau)]$ is defined at any t and τ with $t > \tau$, then all factors on the right hand side of (5.158) are known except $w(\mathbf{x}_0)$. Consequently, the joint probability density function $w(\mathbf{x}_m, \ldots, \mathbf{x}_0)$ of the Markov process, i.e. the Markov process proper, is completely defined by the initial probability density $w(\mathbf{x}_0)$ and the probability density of transitions $w(\mathbf{x}_k, \mathbf{x}_{k-1})$.

A Markov process is said to be a *Gaussian Markov process* if the probability densities for $w(\mathbf{x}_0) = w[\mathbf{x}(t_0)]$ and $w(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ are normal and defined by the respective expectations and dispersion matrices

$$\begin{aligned} \bar{\mathbf{x}}_0 &= E(\mathbf{x}_0) \\ \mathbf{D}_0 &= E[(\mathbf{x}_0 - \bar{\mathbf{x}}_0)(\mathbf{x}_0 - \bar{\mathbf{x}}_0)^{\mathrm{T}}] \\ \bar{\mathbf{x}}_{k|k-1} &= E(\mathbf{x}_k \mid \mathbf{x}_{k-1}) \\ \mathbf{D}_{k|k-1} &= E[(\mathbf{x}_k - \bar{\mathbf{x}}_{k|k-1})(\mathbf{x}_k - \bar{\mathbf{x}}_{k|k-1})^{\mathrm{T}}] \end{aligned} \tag{5.159}$$

5.7.7 Examples of Gaussian Markov Processes

Example 5.17. Let $u(t)$, $t \geqslant 0$, be a scalar process described by the differential equation $\dot{u}(t) = 0$, where $u(0)$ is a Gaussian random variable. Integration yields $u(t) = c = \text{const}$. Consequently, $u(t)$

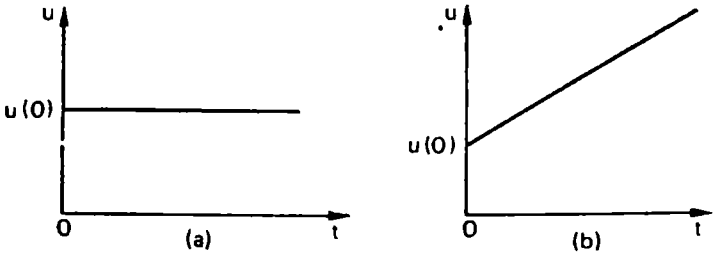


Fig. 5.11. (*a*) A constant signal with a random initial value. (*b*) A signal varying by a linear law

is a constant signal coinciding with $u(0)$ for any $t \geqslant 0$ (Fig. 5.11*a*). Therefore, the value of $u(t)$ for any $t \geqslant 0$ is completely defined by the initial probability density, hence the process is Markov.

Example 5.18. Let $u(t)$, $t \geqslant 0$, be a scalar process described by the second order differential equation $\ddot{u}(t) = 0$ with both $u(0)$ and $\dot{u}(0)$ having a joint normal probability distribution. Integrating this equation yields

$$\dot{u}(t) = \dot{u}(0) = \text{const}$$
$$u(t) = u(0) + t\dot{u}(0)$$

The form of this signal is shown in Fig. 5.11*b*.

Consider the vector quantity

$$\mathbf{x}(t) = \begin{pmatrix} u(t) \\ \dot{u}(t) \end{pmatrix} = \begin{pmatrix} u(0) + t\dot{u}(0) \\ \dot{u}(0) \end{pmatrix}$$

Since $u(0)$ and $\dot{u}(0)$ have a joint normal distribution, then $\mathbf{x}(t)$ is a Gaussian two-dimensional vector. Consequently, the process $\mathbf{x}(t)$ is Gaussian.

To prove that the process is Markov we consider two arbitrary time instants $t_2 > t_1 > 0$. Then

$$\mathbf{x}(t_1) = \begin{pmatrix} \dot{u}(0) + t_1\dot{u}(0) \\ \dot{u}(0) \end{pmatrix}, \quad \mathbf{x}(t_2) = \begin{pmatrix} u(0) + t_2\dot{u}(0) \\ \dot{u}(0) \end{pmatrix}$$

Recognizing that $\dot{u}(0) = \dot{u}(t_1)$ and that $u(0) + t_2\dot{u}(0) = u(0) + t_1\dot{u}(0) + (t_2 - t_1)\dot{u}(0) = u(t_1) + (t_2 - t_1)\dot{u}(t_1)$, we find

$$\mathbf{x}(t_2) = \begin{pmatrix} u(t_1) + (t_2 + t_1)\dot{u}(t_1) \\ \dot{u}(t_1) \end{pmatrix}$$

Consequently, $\mathbf{x}(t_2)$ may be completely defined given $\mathbf{x}(t_1)$. This implies that the process $\mathbf{x}(t)$ is Markov.

Likewise it can be demonstrated that if a scalar process $u(t)$, $t \geqslant 0$, is given by the differential equation $d^{k+1}u(t)/dt^{k+1} = u^{k+1}(t) = 0$ and $u(0), \dot{u}(0), \ldots, u^k(0)$ have a joint normal probability distribution, then the vector process

$$\mathbf{x}(t) = (u(t), \dot{u}(t), \ldots, u^k(t))^{\mathrm{T}}$$

is a normal Markov process.

Example 5.19. White Gaussian noise. Processes in communication and control systems are often corrupted by noise $s(t)$ which is frequently modeled as a white Gaussian noise. Consider a sequence of time instants (index set) $t_1, t_2, \ldots, t_n$ and denote by $s_1, s_2, \ldots, s_n$ the magnitude of noise at these time instants. The interference s_n is called Gaussian noise (white Gaussian sequence) if the values s_n are normally distributed with mean zero, variance σ^2 and do not correlate with one another, namely,

$$\begin{aligned} E(s_n) &= \bar{s}_n = 0 \\ E(s_n^2) &= \sigma^2 \\ E(s_i s_j) &= 0 \quad \text{for} \quad j \neq i \end{aligned} \tag{5.160}$$

As can be seen the magnitude of the noise variable at any time is independent of the previous values of noise and is a normally distributed random variable. Thus, Gaussian white noise is a Gaussian Markov process.

Example 5.20. Correlated normal (Gaussian) noise. The correlated normal noise ξ_n with mean zero $E(\xi_n) = \bar{\xi}_n = 0$, variance $E[\xi_n^2] = \overline{\xi_n^2} = \alpha$, and covariance $\mathrm{cov}_{ij} = E[\xi_i \xi_j] = \alpha\varphi^{|i-j|}$, $0 \leqslant \varphi \leqslant 1$, is a sequence defined by the relation

$$\xi_n = \varphi\xi_{n-1} + s_n, \quad n = 1, 2, \ldots \tag{5.161}$$

where ξ_0 is a random variable distributed to a normal law with mean zero and variance α, and s_n is a Gaussian noise with variance $\sigma^2 = \alpha(1 - \varphi^2)$, uncorrelated with ξ_0.

The fact that the relation (5.161) defines a normal Markov process can be seen from the relation proper and from the definition of ξ_0 and s_n. The above quantities $\bar{\xi}_n$, $\overline{\xi_n^2}$, and cov_{ij} may be readily veri-

fied by considering several leading terms of the sequence ξ_n. For example, for $n = 1$

$$\begin{aligned}
\bar{\xi}_1 &= \varphi\bar{\xi}_0 + \bar{s}_1 \\
\bar{\xi}_1 &= \varphi\bar{\xi}_0 + \bar{s}_1 = 0 \\
\bar{\xi}_1^2 &= E\,[\varphi^2\xi_0^2 + 2\varphi\xi_0 s_1 + s_1^2] = \alpha\varphi^2 + \sigma^2 = \alpha \\
\mathrm{cov}_{01} &= \mathrm{cov}_{10} = E\,[\xi_0\,(\varphi\xi_0 + s_1)] = \alpha\varphi
\end{aligned}$$

and for $n = 2$

$$\xi_2 = \varphi\xi_1 + s_2, \quad \bar{\xi}_2 = 0$$

$$\begin{aligned}
\bar{\xi}_2^2 &= E\,[\varphi^2\xi_1^2 + 2\varphi\xi_1 s_2 + s_2^2] = \alpha\varphi^2 + \sigma^2 = \alpha \\
\mathrm{cov}_{12} &= \mathrm{cov}_{21} = E\,[\xi_1\,(\varphi\xi_1 + s_2)] = \alpha\varphi \\
\mathrm{cov}_{02} &= \mathrm{cov}_{20} = E\,[\xi_0\,(\varphi\xi_1 + s_2)] = \varphi\,\mathrm{cov}_{01} = \alpha\varphi^2
\end{aligned}$$

Noise is frequently represented by a collection of a few sequential values $\xi_1, \ldots, \xi_l$, which may be treated collectively as the noise vector

$$\nu = (\xi_1, \ldots, \xi_l)^{\mathrm{T}} \tag{5.162}$$

characterized by the expectation vector which is a zero vector, and the dispersion matrix $\mathbf{D}_\nu = E\,(\nu\nu^{\mathrm{T}})$

$$\mathbf{D}_\nu = \alpha \begin{pmatrix} 1 & \varphi & \varphi^2 & \ldots & \varphi^{l-1} \\ \varphi & 1 & \varphi & \ldots & \varphi^{l-2} \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \varphi^{l-1} & \varphi^{l-2} & \varphi^{l-3} & \ldots & 1 \end{pmatrix} \tag{5.163}$$

If $\varphi = 0$, then $\xi_n = s_n$, i.e. the correlated noise becomes a white Gaussian noise with variance $\sigma^2 = \alpha$. The dispersion matrix of white Gaussian noise has the form

$$\mathbf{D} = \sigma^2 \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \ldots & 1 \end{pmatrix} = \sigma^2\mathbf{I} \tag{5.164}$$

where $\mathbf{I}$ is the unit matrix.

5.8 Basic Methods of Statistics

5.8.1 Introduction

Many natural phenomena face the researcher as an enormous field of facts and experiments that in turn are influenced by a host of factors. Some of these factors are pertinent to the studied phenomenon, others are of minor importance and often obscure the essence

of the phenomenon. Methods of statistics help the researcher represent his observations in a compact and manageable form devoid of insignificant information. If the available experiments are insignificant and another experiment is required, statistical techniques answer the question how such an experiment should be carried out to obtain the necessary data easy and at low cost. To put long things short, statistics deals with methods of processing of experimental data with the purpose of obtaining useful conclusions.

A variety of problems handled by statistics is detailed in numerous textbooks and monographs [22-25]. Here we limit ourselves by considering the most popular statistical methods. Expansion of some of them to decision making theory will be discussed in Chapter 10.

5.8.2 Random Sample

Let x be a one-dimensional random variable with probability distribution function $F(x)$. Consider the n-dimensional random variate

$$(x^1, \ldots, x^n) \tag{5.165}$$

constituted by independent random variables with identical probability distributions $F(x)$. The cumulative probability distribution function of this multivariate will be defined as the product of the component probability distributions

$$\prod_{i=1}^{n} F(x^{(i)}) \tag{5.166}$$

Random variables may be viewed as realizations of a certain experiment. The multidimensional variate (5.165) may be treated as the result of n independent experiments carried out on the same experimental setup or as the result of n simultaneous experiments on n identical setups.

Because there always exists a possibility, at least theoretical, to perform an unlimited number of experiments, we may speak of an infinite set of random numbers $x^{(i)}$ with probability distribution function $F(x)$. This infinite set is referred to as an *infinite population* with probability distribution $F(x)$.

Each individual result of the experiment, and consequently, each $x^{(i)}$ from the finite set (5.165) may be viewed as a choice of one of the numbers of the population. The entire set (5.165) represents a sequence of n such choices and is called a *random sample*.

5.8.3 Limit Theorems of Probability Theory

In statistics, of particular interest are the properties of a random sample as its size n increases, specifically the limiting relationships as n tends to infinity. Under the circumstances the fundamental role is played by the Chebyshev inequality and its derivations, the law of large numbers and the central limit theorem.

If x is a random variable with mean μ and variance σ^2, then the Chebyshev inequality is

$$P(|x-\mu| \geqslant \lambda\sigma) \leqslant 1/\lambda^2 \tag{5.167}$$

To prove this inequality we divide the real axis into three intervals

$$I = (-\infty,\ \mu - \lambda\sigma]$$
$$I' = (\mu - \lambda\sigma,\ \mu + \lambda\sigma)$$
$$I'' = [\mu + \lambda\sigma,\ +\infty)$$

and represent the variance in the form

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 w(x)\,dx = \int_{I} (x-\mu)^2 w(x)\,dx + \int_{I'} (x-\mu)^2 w(x)\,dx + \int_{I''} (x-\mu)^2 w(x)\,dx$$

If we drop the addend corresponding to the I' interval, then

$$\sigma^2 \geqslant \int_{I} (x-\mu)^2 w(x)\,dx + \int_{I''} (x-\mu)^2 w(x)\,dx$$

Because in the I interval we have $x - \mu \leqslant -\lambda\sigma$ and in the I'' interval $x - \mu \geqslant \lambda\sigma$, the inequality will be only stronger in both intervals if we replace $(x-\mu)^2$ with $\lambda^2\sigma^2$. Thus,

$$\sigma^2 \geqslant \lambda^2\sigma^2 [P(x \leqslant \mu - \lambda\sigma) + P(x \geqslant \mu + \lambda\sigma)]$$
$$= \lambda^2\sigma^2 P(|x-\mu| \geqslant \lambda\sigma)$$

which is equivalent to the inequality (5.167).

The inequality (5.167) defines the probability that the value of $x - \mu$ will be outside the interval $(-\lambda\sigma, \lambda\sigma)$. However, we can also determine the probability that the value of $x - \mu$ will not be outside the said interval. This probability is

$$P(|x-\mu| < \lambda\sigma) = 1 - P(|x-\mu| \geqslant \lambda\sigma)$$

or in view of (5.167)

$$P(|x-\mu| < \lambda\sigma) \geqslant 1 - 1/\lambda^2 \tag{5.168}$$

Denoting $\lambda\sigma$ by ε gives this inequality the form

$$P(|x-\mu|<\varepsilon)\geqslant 1-\sigma^2/\varepsilon^2 \tag{5.169}$$

Consider now a random sample $(x_1, \ldots, x_n)$ of size n. The *sample mean* is

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.170}$$

For brevity we denote the sum over the sample observations by Σ_n so that

$$\bar{x}=\Sigma_n/n=\Sigma(x)/n \tag{5.171}$$

By virtue of (5.62) and (5.98) the variance of $\bar{x}$ will be

$$\operatorname{var}(\bar{x})=\operatorname{var}[\Sigma_n/n]=\frac{1}{n^2}\Sigma_n[\operatorname{var}(x)]=\sigma^2/n \tag{5.172}$$

Substituting in (5.169) $\bar{x}$ for x and σ^2/n for σ^2, we have

$$P\{|\Sigma_n/n-\mu|<\varepsilon\}\geqslant 1-\sigma^2/(n\varepsilon^2) \tag{5.173}$$

where μ is the expectation of $\bar{x}$, i.e. the mean value of the population of which the sample is taken.

From (5.173) it follows that as $n\to\infty$, for any $\varepsilon>0$

$$P\{|\Sigma_n/n-\mu|<\varepsilon\}\to 1 \tag{5.174}$$

This statement represents the *law of large numbers*, saying that at large n, with probability close to unity the *sample mean* coincides with the expectation $\mu=E(x)$ of variable x.

We give without proof one of the more important limit relationships of probability theory, known as the *central limit theorem*. Let $(x_1, \ldots, x_n)$ be a sequence of random variables with generally different probability distribution functions $F_1(x), \ldots, F_n(x)$, means $\mu_1, \ldots, \mu_n$, and variances $\sigma_1^2, \ldots, \sigma_n^2$. Denote

$$\bar{x}=\Sigma(x)/n,\quad \mu=\Sigma(\mu)/n,\quad \text{and}\quad s^2=\Sigma(\sigma^2)/n \tag{5.175}$$

The assertion of the central limit theorem is that if random variables $x_1, \ldots, x_n$ are independent and have finite variances of the same order, then at a large number of addends in the function Σ_n, the distribution of the sample mean $\bar{x}$ approaches the normal distribution with mean μ and variance s^2/n. Observing that in this case the variable

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}} \tag{5.176}$$

has zero mean and unit variance and by the central limit theorem we may assert that as $n \to \infty$

$$P\left(\frac{\bar{x}-\mu}{s/\sqrt{n}} < y\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp(-u^2/2)\, du \qquad (5.177)$$

Problems where the random variable under study is a sum of a large number of independent random variables arise very often in applications. For example, the error of a sophisticated device is a result of a collective action of various external conditions and the errors in the subsystems of the device. Large errors may be easily excluded. For example, an open circuit can be easily traced down and repaired. However, the causes of errors of small significance are hard to evaluate. Consequently, the individual addends of the total error may be deemed uniformly small and we may assert, by the central limit theorem, that the total error will have a distribution close to the normal distribution.

5.8.4 Estimation

One of the problems of mathematical statistics is inference of unknown parameters of the population distribution from a known finite sample. The parameter in question may be the cumulative distribution function $F(x)$ or its parameters such as expectation, variance, the least and largest values of the random variable, and the like.

Because the elements of a random sample are themselves random variables, the value of a parameter determined by this sample will be also random. Specifically, if we have a few samples of the same size taken from the same population, then each of them will give its own value of the desired parameter. Therefore, a finite sample will not provide us with the exact value of a parameter, but rather with a more or less accurate *estimate* of this parameter of the general population.

In general, one and the same sample may be used to obtain different estimates. Let, for example, $x_1, x_2, \ldots, x_n$ be a sample such that $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$ and an estimate of the population mean is desired. Various functions of the elements of the random sample, or *estimators*, can be considered (denoted by an overhead hat), namely,

$$\hat{\bar{x}} = x_{(n+1)/2}, \qquad \text{for odd } n$$
$$= (x_{n/2-1} + x_{n/2+1})/2, \quad \text{for even } n \qquad (5.178)$$

$$\hat{\bar{x}} = (x_1 + x_n)/2 \qquad (5.179)$$

$$\hat{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (5.180)$$

Which of these estimators is preferable? To answer this question we need certain criteria which the estimator should satisfy. Three most important of them are given below.

Unbiasedness. An estimator should not involve a systematic error that would increase or decrease the value of a parameter for all samples. This requirement implies that the expectation of an estimator must coincide with the actual value of the estimated parameter. If we denote the actual value of a parameter by θ and its estimator by $\hat{\theta}$, then the requirement that the estimator for θ should be unbiased will be

$$E(\hat{\theta}) = \theta \tag{5.181}$$

Accordingly, the difference $b(\theta) = E(\hat{\theta}) - \theta$ is called the *bias* of θ.

Consistency. The estimate $\hat{\theta}$ must approach the value of parameter θ for larger sample sizes. Since the estimator $\hat{\theta}$ is a random variable, we may say of this approach in the probability sense only. If we denote by $\hat{\theta}_n$ an estimator of θ obtained from a sample of size n, then this estimator will be *consistent* if for any $\varepsilon > 0$

$$P(|\hat{\theta}_n - \theta| < \varepsilon) \to 1 \tag{5.182}$$

as $n \to \infty$.

Efficiency. Of all unbiased and consistent estimates based on the same sample, the one should be preferred that appears closer to the estimated parameter. This degree of closeness may be judged by a smaller number of large deviations, i.e. by the variance. The relative efficiency of two unbiased estimates $\hat{\theta}'_n$ and $\hat{\theta}''_n$ of a parameter θ may be defined as $\text{var}(\hat{\theta}'_n)/\text{var}(\hat{\theta}''_n)$, the efficiency of $\hat{\theta}'_n$ relative to $\hat{\theta}''_n$ being greater than unity if $\text{var}(\hat{\theta}'_n) < \text{var}(\hat{\theta}''_n)$. Therefore, an efficient estimate is one of minimum variance, i.e.

$$\text{var}(\hat{\theta}) = \min \tag{5.183}$$

5.8.5 Unbiased Estimates of Expected Value and Variance

Let $x_1, \ldots, x_n$ be a sample of a population with expected value μ and variance σ^2. Denote by $\bar{x}$ an estimate for the expectation and by s^2 an estimate for the variance.

The estimator of μ is usually the formula for the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum (x) \tag{5.184}$$

From the law of large numbers it follows that this estimator is consistent, i.e. approaches μ as n tends to infinity.

Because the sample mean $\bar{x}$ is the sum of random variables, it is itself a random variable. Therefore, we may speak of a *sample mean distribution* and of its expected value and variance.

The expected value of a sample mean

$$E(\bar{x}) = \frac{1}{n} E[\Sigma(x)] = \frac{1}{n} \Sigma[E(x)] = \mu \tag{5.185}$$

coincides with the expected value of the population (population mean), that is, the estimator (5.184) is unbiased. According to (5.172) the variance of the sample mean is σ^2/n.

If the sample is taken from a population with normal distribution of the random variable, then $\bar{x}$ will be distributed normally too as a sum of normal random variables. By virtue of (5.185) and (5.172) we write this distribution as

$$w(\bar{x}) = N(\mu, \sigma^2/n) \tag{5.186}$$

In many applications it is more convenient to consider in place of $\bar{x}$ the variable

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \tag{5.187}$$

which has zero mean and unit variance; its probability density is denoted by $N(0, 1)$.

If a sample is taken from a population other than normal, the distribution of the sample mean will also differ from normal, but at large sample sizes it will be close to $N(\mu, \sigma^2/n)$ on the base of the central limit theorem.

To evaluate the unbiased estimate s^2 for the variance we find the expected value of $\Sigma(x - \bar{x})^2$. We write this quantity in the form

$$\begin{aligned}\Sigma(x - \bar{x})^2 &= \Sigma[(x - \mu) - (\bar{x} - \mu)]^2 \\ &= \Sigma(x - \mu)^2 - 2(\bar{x} - \mu)\Sigma(x - \mu) + n(\bar{x} - \mu)^2\end{aligned}$$

and by taking the expected value of each addend we obtain

$$E[\Sigma(x - \bar{x})^2] = (n - 1)\sigma^2 \tag{5.188}$$

or

$$E[\Sigma(x - \bar{x})^2/(n - 1)] = \sigma^2 \tag{5.189}$$

From (5.189) it follows that an unbiased estimator of the variance may be the following

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 \tag{5.190}$$

Before we determine the distribution of the estimate of the variance, we consider an important distribution law. If y_i are independent random variables obeying a normal distribution N (0, 1), then the random variable

$$\chi^2 = \sum_{i=1}^{n} y_i^2 \tag{5.191}$$

obeys the χ^2 distribution (read chi-square) with n degrees of freedom. The number of degrees of freedom is defined by the number of independent random variables in the sum (5.191). The χ^2 distribution is tabulated for various n as the probabilities $p\,(\chi^2 > \chi_\alpha^2)$, where χ_α^2 is a certain positive number. Table 5.2 given below is an example of such a table for n from 1 to 30 and χ_α^2 satisfying the condition $p\,(\chi^2 > \chi_\alpha^2) = 0.95$.

Table 5.2. Chi-square Distribution Function for $\alpha = 0.05$

n	χ_α^2	n	χ_α^2	n	χ_α^2	n	χ_α^2	n	χ_α^2	n	χ_α^2
1	.004	6	1.635	11	4.575	16	7.962	21	11.591	26	15.379
2	.103	7	2.167	12	5.226	17	8.682	22	12.338	27	16.151
3	.352	8	2.733	13	5.892	18	9.390	23	13.090	28	16.928
4	.710	9	3.325	14	6.571	19	10.117	24	13.848	29	17.708
5	1.145	10	3.940	15	7.261	20	10.851	25	14.611	30	18.493

A more elaborated table of chi-square distribution (for α other than 0.05) may be found in any text that is completely devoted to mathematical statistics, e.g. [23].

It is not hard to see that if y_i is taken as $(x_i - \bar{x})/\sigma$ satisfying the conditions imposed on y_i, then the function

$$\chi^2 = \sum_{i=1}^{n} y_i^2 = \frac{(n-1)\,s^2}{\sigma^2} \tag{5.192}$$

will give us the distribution law for s^2.

5.8.6 Maximum Likelihood Estimation

Many problems of mathematical statistics reduce to estimation of a parameter θ in a distribution of known type. The method of maximum likelihood is a method of estimation based on a different rationale than is the method of moment discussed above. Rather than equating sample moments to population moments, it sets the

estimate equal to the value of the unknown parameter which maximizes the probability of the observed sample results. This, in a sense, gives that estimate which is the "most likely" value for the unknown parameter, based upon the observed sample.

Suppose that we have a sample $x_1, \ldots, x_n$ of a population with probability density function $w(x, \theta)$ depending upon a parameter θ. The multidimensional probability density function

$$L = \prod_{i=1}^{n} w(x_i, \theta) \tag{5.193}$$

for this sample is called the *likelihood function*. According to the principle of maximum likelihood, the estimate of a parameter θ is the value $\hat{\theta}$ that maximizes L or $\log L$, since the log function is monotonic increasing. The value $\hat{\theta}$ can be found from the equation

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log w(x_i, \theta)}{\partial \theta} = 0 \tag{5.194}$$

which can easily be reduced to the form

$$\sum_{i=1}^{n} \frac{1}{w(x_i, \theta)} \frac{\partial w(x_i, \theta)}{\partial \theta} = 0 \tag{5.195}$$

For the case when the sample is taken from a normal population $N(\mu, \sigma^2)$ with probability density given by (5.40), the likelihood equations solved for parameters μ and σ are as follows

$$\sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\sum_{i=1}^{n} \left[1 - \frac{(x_i - \mu)^2}{\sigma^2}\right] = 0 \tag{5.196}$$

The first equation yields the estimate for μ

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \tag{5.197}$$

which coincides with the unbiased and consistent estimate (5.184) derived above. The second equation yields an estimate for σ^2

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5.198}$$

which, however, is not unbiased. Therefore, for practical purposes the unbiased estimate (5.190) is preferred.

An example below illustrates the application of the method of maximum likelihood to deriving an estimate from an exponential distribution.

Example 5.21. To evaluate the average life expectancy of equipment, a sample of n devices was tested for t hours. Of this sample, m devices survived the test, and the other $n - m$ units failed at times $\tau_1, \ldots, \tau_{n-m}$. What is the most likely estimate of the mean time to failure (MTTF)?

If we assume that the time to failure obeys the exponential law (5.135), then the probability of failure at time τ is $w(\lambda, \tau) = \lambda \exp(-\lambda\tau)$, and the probability that the device will not fail for time t is $\exp(-\lambda t)$. The likelihood function is

$$L = (e^{-\lambda t})^m \prod_{i=1}^{n-m} w(\lambda, \tau_i) \tag{5.199}$$

Equation (5.194) takes the form

$$-mt + \sum_{i=1}^{n-m} \left(\frac{1}{\lambda} - \tau_i\right) = 0 \tag{5.200}$$

whence

$$\text{MTTF} = \frac{1}{\lambda} = \left(mt + \sum_{i=1}^{n-m} \tau_i\right) \Big/ (n-m) \tag{5.201}$$

In a similar manner we can demonstrate that the most likely estimate of the probability of an event for the independent trials process obeying a binomial distribution is

$$\hat{p} = x/n \tag{5.202}$$

where n is the number of trials, and x is the number of favorable events (successes).

If the random process obeys a Poisson distribution and a sample of n intervals contains r events, then the most likely estimate of the average number of events per interval is

$$\hat{a} = r/n \tag{5.203}$$

5.8.7 Confidence Interval Estimation

In the preceding sections we discussed methods that provide an estimate of a parameter in the form of a constant, i.e. some point on the real axis. These estimates are therefore referred to as point estimates. Because a sample is taken at random, in many cases,

especially at small sample sizes, this estimate may appear rather far from the actual value of this parameter. If we use a point estimator of an unknown parameter, we gain some idea about the numerical value of the parameter. But, if we do nothing more, we have no idea how confident we might be, given our particular sample, that our guess is anywhere close to the actual value of the parameter.

In this section we shall discuss confidence intervals for parameters; these provide both an idea of the numerical magnitude of the parameter and of how confident we may be that a given interval includes the true unknown parameter value. To be confident, we choose a probability, called *confidence coefficient* and denoted $1 - \alpha$, close to unity by taking α, the *confidence level*, sufficiently small, most commonly $\alpha = 0.05$ or 0.01. If L_1 and L_2 are statistics such that

$$P(L_1 \leqslant \theta \leqslant L_2) \geqslant 1 - \alpha \tag{5.204}$$

for all possible values of an unknown parameter θ, then the interval (L_1, L_2) is a $100(1 - \alpha)\%$ *confidence interval* for θ at the confidence level α. The end points l_1 and l_2 (observations of L_1 and L_2) of the confidence interval are called *confidence limits.*

Now given one particular sample and the observed values l_1 and l_2 we have no idea whether we have in fact bracketed θ; clearly we either do bracket it or we do not. But, since the two random variables L_1 and L_2 will bracket θ with probability at least $1 - \alpha$, we say that we are at least $100(1 - \alpha)\%$ confident that this particular sample has led to an interval that brackets θ.

If $\hat{\theta}$ is a known point estimate of parameter θ and the probability density function $w(\hat{\theta})$ for this estimate is also known, then we can easily find a central interval (L_1, L_2) containing a fraction $p = 1 - \alpha$ of the population. The end points L_1 and L_2 satisfy the relation

$$\int_{L_1}^{L_2} w(\hat{\theta})\, d\hat{\theta} = 1 - \alpha \tag{5.205}$$

Some problems require that in place of a two-sided confidence interval lower or upper bounds should be found for the parameter in question θ. The intervals defined as

$$\begin{aligned} P(\theta > l_1) &= 1 - \alpha \\ P(\theta < l_2) &= 1 - \alpha \end{aligned} \tag{5.206}$$

are referred to as *one-sided* $100(1 - \alpha)\%$ *confidence intervals.* The limits of these intervals l_1 and l_2 obey the relations

$$\int_{l_2}^{\infty} w(\hat{\theta})\, d\hat{\theta} = 1 - \alpha \quad \text{and} \quad \int_{-\infty}^{l_2} w(\hat{\theta})\, d\hat{\theta} = 1 - \alpha \tag{5.207}$$

Let us assume that a sample of size n is taken from a population with normal p.d.f. $N(\mu, \sigma^2)$ of unknown μ and known σ^2, so that the variable $t = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ will be distributed with mean zero and unit variance, or $N(0, 1)$. In this case the $100(1-\alpha)$ % confidence interval will be defined as

$$P(-t_\alpha < \sqrt{n}(\bar{x} - \mu)/\sigma < t_\alpha) = 1 - \alpha \tag{5.208}$$

or

$$P(\bar{x} - t_\alpha \sigma/\sqrt{n} < \mu < \bar{x} + t_\alpha \sigma/\sqrt{n}) = 1 - \alpha \tag{5.209}$$

By virtue of (5.44) this expression can be rewritten as

$$1 - \alpha = \Phi(t_\alpha) - \Phi(-t_\alpha) = 2\Phi(t_\alpha) \tag{5.210}$$

This expression leads to the following formula for t_α

$$\Phi(t_\alpha) = (1 - \alpha)/2 \tag{5.211}$$

For one-sided confidence intervals (t'_α, ∞) and $(-\infty, t'_\alpha)$ we put (5.210) to the form

$$1 - \alpha = \Phi(t'_\alpha) - \Phi(-\infty) = \frac{1}{2} + \Phi(t'_\alpha) \tag{5.212}$$

For the end values of one-sided confidence intervals we get

$$\Phi(t'_\alpha) = \frac{1}{2} - \alpha \tag{5.213}$$

Thus, the limits of the $100(1 - \alpha)$% one-sided confident interval are equal to that of the $100(1 - 2\alpha)$% two-sided confidence interval.

If we let $\alpha = 0.05$ and make use of the tables of the standard normal integral, then from (5.211) and (5.213) we obtain the following limits for two-sided and one-sided confidence intervals

$$t_\alpha = 1.96 \quad \text{and} \quad t'_\alpha = 1.64 \tag{5.214}$$

Thus, if we take a sample of size n from a normal population of variance σ^2 and compute the sample mean $\bar{x}$, then we may assert that with probability 0.95 the expected value of the population will lie in the interval

$$\bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n} \tag{5.215}$$

or

$$\mu > \bar{x} - 1.64\sigma/\sqrt{n} \tag{5.216}$$

or

$$\mu < \bar{x} + 1.64\sigma/\sqrt{n} \tag{5.217}$$

These values of t_α may be used for the evaluation of confidence intervals under binomial distribution $w(x, n, p)$ since at large n the binomial distribution approaches the normal distribution with mean $\mu = np$ and variance $\sigma^2 = npq$. By letting $t = (x - np)/\sqrt{npq}$

we conclude that if a favorable event has occurred x times in n trials, then with probability 0.95 the expected value np will lie in the following confidence limits

$$x - 1.96\sqrt{npq} < np < x + 1.96\sqrt{npq} \tag{5.218}$$

or

$$np > x - 1.64\sqrt{npq} \tag{5.219}$$

or

$$np < x + 1.64\sqrt{npq} \tag{5.220}$$

The value $\sqrt{npq}$ is calculated by the most likely value of $\hat{p} = x/n$ as

$$\sqrt{npq} = \sqrt{x(n-x)/n} \tag{5.221}$$

Example 5.22. A batch of 100 items sampled at random from a large lot of manufactured goods was inspected to reveal a certain type of defect. The number of defected parts was 37. What can we say about the percentage of defected items in the whole lot?

In this case, $n = 100$, $x = 37$, $\sqrt{npq} = 4.94$, so that $1.96 \times \sqrt{npq} = 9.5$ and

$$x - 1.96\sqrt{npq} = 27.5$$

$$x + 1.96\sqrt{npq} = 46.5$$

Thus, there is a chance of 5 in 100 that we shall be in error by making one of the assertions about the proportion of items having defects as follows: (1) between 27.5% and 46.5% of the whole lot, (2) at least, 29% of the whole lot, and (3) at most 45% of the whole lot.

Confidence intervals are often used in estimations of the mean μ under the normal distribution whose variance σ^2 is unknown. In this case we estimate the variance by s^2. The random variable

$$t = (\bar{x} - \mu)/\sqrt{s^2/n} \tag{5.222}$$

is not normal any longer, but obeys the Student t distribution defined as follows.

Let u be a random variable with normal distribution $N(0, 1)$ and v be a random variable with the chi-square distribution. If u and v are statistically independent, then the random variable

$$t = u/\sqrt{v/\nu} = u\sqrt{\nu}/\sqrt{v} \tag{5.223}$$

has Student's t distribution on ν degrees of freedom.

If we define the variables as

$$u = (\bar{x} - \mu)/\sqrt{\sigma^2/n} \quad \text{and} \quad v = (n-1)s^2/\sigma^2 \tag{5.224}$$

then by (5.187) and (5.192) they will satisfy the conditions defining Student's t distribution, and the value of t given by (5.223) coincides with that obtained by (5.222) at $\nu = n - 1$. Thus, the random variable t given by (5.222) will have Student's t distribution on $n - 1$ degrees of freedom.

Values of $t_{\alpha\nu}$, the limits of a two-sided 100 $(1 - \alpha)$% confidence interval for Student's t distribution with $n - 1$ degrees of freedom, are tabulated (see, e.g. [23]) for various α and ν from 1 to 30. For $\nu > 30$ the values of $t_{\alpha\nu}$ practically coincide with the values of t_α obtained for the normal distribution. By way of example Table 5.3 presents the values of $t_{\alpha\nu}$ for $\alpha = 0.05$ and 0.10.

Table 5.3. Distribution of t

DF \ α	0.05	0.10	DF \ α	0.05	0.10	DF \ α	0.05	0.10
1	12.71	6.31	11	2.20	1.80	21	2.08	1.72
2	4.30	2.92	12	2.18	1.78	22	2.07	1.72
3	3.18	2.35	13	2.16	1.77	23	2.07	1.71
4	2.78	2.13	14	2.15	1.76	24	2.06	1.71
5	2.57	2.01	15	2.13	1.75	25	2.06	1.71
6	2.45	1.94	16	2.12	1.75	26	2.06	1.71
7	2.36	1.89	17	2.11	1.74	27	2.05	1.71
8	2.31	1.86	18	2.10	1.73	28	2.05	1.70
9	2.26	1.83	19	2.09	1.73	29	2.05	1.70
10	2.23	1.81	20	2.09	1.73	30	2.04	1.70
						∞	1.96	1.64

The limits of one-sided confidence intervals $t'_{\alpha\nu}$ can be found by the values of $t_{\alpha\nu}$ observing that

$$t'_{\alpha\nu} = t_{2\alpha\nu} \tag{5.225}$$

Example 5.23. To reduce the effect of noise in the communication system, each measurement on board a spacecraft is sent to the ground station in three separate messages. Assume that the results of the measurements were $x_1 = 3.2$, $x_2 = 2.9$, and $x_3 = 3.1$. Determine the two-sided 95% confidence interval for the measured parameter, assuming that the distortions of the signal in each message are statistically independent and distributed to a normal law.

By the measurement data we find $n = 3$, $\bar{x} = \Sigma\,(x)/n = 3.067$, and $x^2 = \Sigma\,(x - \bar{x})^2/(n - 1) = 0.0234$.

From Table 5.3 we find for $\alpha = 0.05$ and DF $= n - 1 = 2$ that $t_{\alpha\nu} = 4.30$, so that $t_{\alpha\nu}s/\sqrt{n} = 0.377$. Consequently, with probability 0.95 our value of the parameter may be asserted to lie between 2.69 and 3.44.

5.8.8 Tests of Hypotheses

Let Z be the sample space of an experiment with elements $z \in Z$ To evaluate the probability $p(z)$ of a certain outcome z we carry out n experiments. If the outcome z occurs n_z times, then the probability of this outcome will be estimated by the ratio n_z/n. In fact, however, the ratio n_z/n gives the frequency $q(z)$ rather than probability $p(z)$, which may significantly differ from the probability, especially if the number of experiments is not very large. It is interesting, therefore, to investigate how reliable are our guesses on the probability distribution of the outcomes, $p(z)$, based on the frequencies of these outcomes $q(z)$. However, this problem statement is not entirely correct. Indeed, if the experiment results in a number of figures that express the frequencies of some outcomes $q(z)$, then to replace them with another set of numbers (probabilities $p(z)$) some grounds are needed. What can serve as a ground for replacing one numbers with others?

Assume that in 10 throwings of a coin we have 7 heads. Then the frequency of heads in this experiment is 0.7. Do we have sufficient grounds to replace this value and take the probabilities of heads to some other value? If we know nothing about the coin, i.e. whether it is symmetric or not, we have no grounds to change the frequency value obtained in the experiment.

The problem, however, may be stated in another way. We know that when a coin is symmetric, the probability of heads is 0.5. We wish to check whether or not it is symmetric and toss it 10 times. Heads come up seven times. Now the question is what are our grounds to assert that the coin is symmetric?

Formulated in this way, the problem is called a *test of hypothesis*. Problems of these type arise where there are some grounds to construct some hypothesis (say, the coin is symmetric) about the nature of the distribution or the parameters of the distribution of a random variable. The objective of such an experiment is to prove or disprove the hypothesized statement. The methods of mathematical statistics allow the statistitian to answer, for example, the following questions: (1) will the novel device demonstrate better performance than the existent counterparts? (2) will the system possess a reliability not below a certain level? (3) will the new medicine be more efficacious than the existent methods?

In order to have certain yardstick for acceptance or rejection of a hypothesis in question, we need to develop some criterion that will check that the hypothesis is consistent with the results of the experiment.

Let us assume that the value of some parameter θ is θ_0 and test whether or not this hypothesis is true experimentally. The experiment yields an estimate of this parameter $\hat{\theta}$ which, being a random variable, does not in general coincide with θ_0. The deviation of the estimate $\hat{\theta}$ from the given value θ_0 should not be large, and if this deviation $\Delta\theta$ is found to be large, then it cannot be explained by casual reasons and the hypothesis on the parameter being equal to θ_0 should be rejected. Quite appropriately, the choice of a criterion has reduced to specifying a value of critical deviation $\Delta\theta_{cr}$ such that the probability of exceeding this value will be very small.

Most often than not, the critical deviation is taken to be the limit of a 100 $(1 - \alpha)$% confidence interval. In such a setting, we assert that there is a probability α that the observed deviation exceeds the critical region and the hypothesis will be rejected which in fact is true (*type I error*). The probability called the *significance level of the test*, should be selected depending on those consequences to which the rejection of the true hypothesis might lead. For most situations $\alpha = 0.05$ is a good choice, however, when the false rejection of the true hypothesis may entail severe consequences, smaller values of type I error probability are assumed, such as $\alpha = 0.01$ or even smaller.

It is instructive to note that although this method allows the statistitian to reject false hypotheses with high confidence, it cannot serve as a proof that the hypothesis is true. Accordingly, when a hypothesized value θ_0 appears to lie in the confidence interval, we accept the hypothesis as consistent with the results of the experiment, rather than as a true hypothesis. A separate investigation is needed to judge on whether or not the particular hypothesis is true.

5.8.9 Estimation of the Effect of a Factor on a Random Variable

A frequent problem is to answer the question of whether or not a parameter A influences the behavior of a random variable y. Problems of this type include estimation of efficacy of a new medical treatment, effect of a new treatment of land on the harvest, effect of alterations in technology on the percentage of rejects, and the like.

To answer the above question, the analyst performs n experiments to determine in each of them the values of y_i' and y_i'' with the factor A on and off, respectively, and calculate the difference $x_i = y_i' - y_i''$.

Assume that the random variable in question is normal $N(\mu, \sigma^2)$ with variance σ^2, and that the effect of factor A reduces to a change in the mean μ. If the effect of A is not significant, then the expectations of random variables y' and y'' will be identical and the difference $x = y' - y''$ will be normally distributed with mean zero and variance $\sigma_x^2 = 2\sigma^2$. Therefore, the answer to the above question boils down to testing the hypothesis on the expectation of $x = y' - y''$ being zero.

The sample mean of n tests is

$$\bar{x} = \frac{1}{n} \Sigma (x) = \frac{1}{n} \Sigma (y' - y'') \tag{5.226}$$

If the hypothesis is true, then this variable will be normal with mean zero and sample variance σ_x^2/n. Using the 95% confidence interval we may assert that the hypothesis of insignificance of factor A does not contradict the experimental data if

$$|\bar{x}| < 1.96 \sqrt{\sigma_x^2/n} \tag{5.227}$$

However, in many a situation the variance σ^2 is unknown and should be estimated by s_x^2 calculated by the same experimental data with the formula (5.190).

The quantity $\bar{x}/\sqrt{s^2/n}$ is a random variable obeying the Student t distribution with $n - 1$ degrees of freedom. Accordingly, the hypothesis on factor A being insignificant is accepted if

$$\left| \frac{\bar{x}}{\sqrt{s^2/n}} \right| < t_\alpha \tag{5.228}$$

where t_α is the limit of the 95% confidence interval for Student's t distribution with $n - 1$ degrees of freedom, that may be looked up in, say, Table 5.3.

5.8.10 Variance Hypothesis Testing

In some problems it is desired to test equality of the variances of two independent samples $(x_1, \ldots, x_{n_1})$ and $(y_1, \ldots, y_{n_2})$. The variances of these samples s_1^2 and s_2^2 are calculated by the formula (5.190) with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively. To test whether these samples are taken from the same population with variance σ^2 or from different populations with variances σ_1^2 and σ_2^2 the hypothesis of equality of these variances is tested by the F test in which the F-ratio is computed

$$F = s_1^2/s_2^2 \tag{5.229}$$

where it is assumed that $s_1^2 > s_2^2$ and the hypothesis about the samples being taken from different populations is accepted if the values of F ratio exceed some critical value F_α. For the given degree of freedom, the critical value F_α will be one for which $P(F > F_\alpha) = \alpha$. The value F_α is the upper limit of F distribution corresponding to the probability α. The distribution of F is tabulated usually for two probabilities $\alpha = 0.05$ and $\alpha = 0.01$ for degrees of freedom DF_1 from 1 to 500 and DF_2 from 1 to 1000 in fundamental texts on statistics and in mathematical handbooks (see, e.g. [23]).

5.9 Regression Analysis and Experimental Design

5.9.1 The Problem of Regression Analysis

Suppose that we have k variables $x_1, \ldots, x_k$ and a quantity y depending upon them. The variables are not necessarily random and, if desired, may be specified at will. However, the quantity y is influenced by other factors escaping accurate control or observation, so that y is a random variable. Here we are interested in experimental methods of evaluating the effect of variables $x_1, \ldots, x_k$ on y [26].

Example 5.24. This type of problems is usually connected with the analysis of technological conditions for industrial or laboratory processes. An example may be the process of synthesis of ammonia by removing it from the gases evolving in the cooking process. These gases contain nitrogen, hydrogen, and carbon dioxide and should be freed from the latter. Then the synthesis of ammonia occurs by the scheme

$$3H_2 + N_2 = 2NH_3$$

The purification of gases occurs by passing them through a layer of moving water that absorbs the carbon dioxide. Let the quality of the purification be characterized by the residual percentage y of carbon dioxide in the purified gas. This parameter if affected first of all by the temperature of the gases and the flow rate of water, denoted by x_1 and x_2. However, the quality of purification is also affected by a host of random factors, beginning from the quality of the coal and terminating with the ambient temperature.

Fig. 5.12. Block diagram of a technological process

Let us assume that $x_1, \ldots, x_k$, commonly called *factors*, are input variables of a technological process and y is the output variable of this process, as shown in Fig. 5.12. The y variable will consist of a deterministic component $f(x_1, \ldots, x_k)$ caused by the effect

and variability of the factors $x_1, \ldots, x_k$ and a component ξ due to the action of random factors so that

$$y = f(x_1, \ldots, x_k) + \xi \tag{5.230}$$

The component ξ will be deemed a random variable having a normal distribution with mean zero. In this case the deterministic component will be a conditional expectation of y given the values of the factors $x_1, \ldots, x_k$, namely,

$$f(x_1, \ldots, x_k) = E(y \mid x_1, \ldots, x_k) = \bar{y}(x_1, \ldots, x_k) = \bar{y}(\mathbf{x}) \tag{5.231}$$

Here, $\mathbf{x} = (x_1, \ldots, x_k)^T$ is the vector of values of the input variables that may be viewed as a point in the k-dimensional space of the variables $x_1, \ldots, x_k$, called the *factor space.*

The main purpose of our further consideration is to determine by the experimental data the type of dependence (5.231) which represents some surface in factor space referred to as a *response surface.*

For the purpose of the further analysis it would be convenient to expand the function $f(x_1, \ldots, x_k)$ in a Taylor series about a point of interest and take a finite number of terms, so that a representation of the response function as a polynomial of finite order results

$$\bar{y}(\mathbf{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{j=1,\, i>j}^{k} \beta_{ij} x_i x_j + \ldots \tag{5.232}$$

This expression may be represented in a more compact form by introducing a dumb variable $x_0 = 1$ and denoting the products of variables by the scheme $x_1^2 = x_{k+1}$, $x_2^2 = x_{k+2}$, $\ldots$, $x_k^2 = x_{2k}$, $x_1 x_2 = x_{2k+1}$, $x_1 x_3 = x_{2k+2}$, $\ldots$. This yields the total number of variables to

$$k' = \binom{k+d}{k}$$

where k is the initial number of independent variables, d is the degree of polynomial representing the function $\bar{y}(\mathbf{x})$. In this new notation the equation of the response surface becomes

$$\bar{y}(\mathbf{x}) = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} x_{k+1} + \ldots + \beta_{k'} x_{k'} = \sum_{i=0}^{k'} \beta_i x_i \tag{5.233}$$

For simplicity we shall drop primes of the indexes assuming that the response surface is given by equation (5.233) in k variables. This equation is known as the *regression equation* and the coefficients β_i are called *regression coefficients.* The object of *regression analysis* consists in experimental evaluation of regression coefficients by

observing the behavior of the input variables $x_1, \ldots, x_k$ and dependent (output) variable y. This aim may be achieved by the methods of passive and active experimentation.

Passive experimentation consists in recording the controlled parameters in normal operation of the process without having to introduce any prearranged disturbances. Active experimentation involves predetermined perturbations introduced into the process according to a schedule arranged in advance. Each of these methods has its own advantages and disadvantages.

Artificial perturbations allow the experimenter to rapidly reveal the desired relationships between the parameters and localize the regions of optimal performance. It should not be overlooked, however, that artificial perturbations may lead to distortion of the normal process behavior.

In a passive experiment, the experimenter simply records the natural manifestations of the desired relationships. This approach takes much more experimental time than the active experiment and results in a mathematical description for the region close to the operating point which may lie far from the optimal region.

When the number of experiments is limited, the probability is high that the accurate values of the regression coefficients β_i would not be found, and the experimenter has to employ the *estimates* $\hat{\beta}_i = b_i$ of these coefficients, and the equation of regression becomes

$$\hat{y} = \sum_{i=0}^{k} b_i x_i \tag{5.234}$$

It defines the estimate of the mean $\hat{y}$ (**x**) rather than the expected value $\bar{y}$ (**x**) itself.

5.9.2 Evaluation of Regression Coefficients

Equation (5.234) has $k + 1$ unknown coefficients of regression. To determine these coefficients, a number of experiments are carried out in each of which all variables at the input and output of the process are measured. The total number of experiments n should not be less than the number of the unknown regression coefficients.

Consider an experiment labeled l. Denote by $x_i^{(l)}$, $i = 0, \ldots, k$, and y_l the values of x_i and y observed in this experiment. The estimate

$$\hat{y}_l = \sum_{i=0}^{k} b_i x_i^{(l)}, \quad l = 1, \ldots, n \tag{5.235}$$

determined by the estimator (5.234) will differ from the actually

measured value y_l. The difference between the measured value y_l and its estimate $\hat{y_l}$ will be referred to as a *residual*

$$\varepsilon_l = y_l - \hat{y_l} = y_l - \sum_{i=0}^{k} b_i x_i^{(l)}, \quad l = 1, \ldots, n \tag{5.236}$$

or now the dependent variable can be written as

$$y_l = \hat{y_l} + \varepsilon_l = \sum_{i=0}^{k} b_i x_i^{(l)} + \varepsilon_l, \quad l = 1, \ldots, n \tag{5.237}$$

Unlike the quantity ξ in (5.230) which is purely random, the residual ε_l in (5.236) depends on the values of the regression coefficients b_i. Therefore b_i may be determined by the method of least squares. According to this method, the best estimates b_i will be those which minimize the sum of squares of the residuals,

$$Q = \sum_{l=1}^{n} \varepsilon_l^2 = \min \tag{5.238}$$

To make our further consideration neater we shall use the vector-matrix notation. Denote by $\mathbf{B} = (b_0, b_1, \ldots, b_k)^{\mathrm{T}}$ the column vector of the unknown regression coefficients, and by $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ the column vector of measurements of the conditioned variable y. The observations of the *predictor variables* $x_1, \ldots, x_k$ will be arrayed in the matrix of experimental data

$$\mathbf{X} = \begin{pmatrix} x_0^{(1)} & \ldots & x_k^{(1)} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_0^{(n)} & \ldots & x_k^{(n)} \end{pmatrix} \tag{5.239}$$

The column vector of residuals will be denoted by $\mathbf{E} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$.

In this notation equation (5.237) rewrites as

$$\mathbf{Y} = \mathbf{BX} + \mathbf{E} \tag{5.240}$$

whence

$$\mathbf{E} = \mathbf{Y} - \mathbf{XB} \tag{5.241}$$

The sum of squares of residuals in matrix form is

$$Q = \sum_{l=1}^{n} \varepsilon_l^2 = \mathbf{E}^{\mathrm{T}}\mathbf{E} = (\mathbf{Y} - \mathbf{XB})^{\mathrm{T}} (\mathbf{Y} - \mathbf{XB}) \tag{5.242}$$

or

$$Q = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{XB} - \mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XB} \tag{5.243}$$

It is easy to note that the third addend of this expression equals the transposed second addend. Since Q is a scalar, the terms of this sum are also scalars which do not change under transposition, hence

the second and third addends may be combined so that (5.243) becomes

$$Q = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{B}^T\mathbf{X}^T\mathbf{Y} + \mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{B} \tag{5.244}$$

An equation for $\mathbf{B}$ can be obtained by taking the vector derivative of Q with respect to $\mathbf{B}$ and equating it to zero. Making use of the rule for vector differentiation (see Sect. 4.7), we have

$$\partial Q/\partial \mathbf{B} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\mathbf{B} = 0 \tag{5.245}$$

whence

$$\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{5.246}$$

All the matrix computations involved in (5.246) are usually relegated to a computer. However, in simple situations we can arrive at the analytical expressions by making these calculations by hand.

Example 5.25. Consider a plant with one-dimensional input (single predictor variable) $x_1 = x$, for which the regression equation has the form

$$y = b_0 + b_1 x$$

so that

$$\mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x^{(1)} \\ \cdot & \\ \cdot & \\ 1 & x^{(n)} \end{pmatrix}$$

and

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_l x^{(l)} \\ \sum_l x^{(l)} & \sum_l (x^{(l)})^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_l y_l \\ \sum_l x^{(l)} y_l \end{pmatrix} = n \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}$$

As usual, an overhead bar indicates the mean value of the respective variables defined by (5.184 .

Recognizing that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n(\overline{x^2} - \bar{x}^2)} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

from (5.246) it follows that

$$\mathbf{B} = \frac{1}{\overline{x^2} - \bar{x}^2} \begin{pmatrix} \overline{x^2}\bar{y} - \bar{x}\,\overline{xy} \\ -\bar{x}\,\bar{y} + \overline{xy} \end{pmatrix}$$

so that

$$b_0 = (\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy})/(\overline{x^2} - \overline{x}^2)$$
$$b_1 = (\overline{x}\,\overline{y} - \overline{x}\,\overline{y})/(\overline{x^2} - \overline{x}^2)$$

In actual situations the estimation of the regression coefficients $b_0, b_1, \ldots, b_k$ is carried out by experimental data corrupted by noise, therefore to obtain accurate estimates the number of experiments n should esceed considerably the number $k + 1$ of the estimated regression coefficients. The difference between the number of observations and the number of the unknown parameters estimated by these observations is called the *degrees of freedom of the experiment.* In regression analysis this number of degrees of freedom is $n' = n - (k + 1)$.

Another source of errors may be inconsistency between the degree of the polynomial describing the response surface and the actual character of this surface. Whether or not a choice of a polynomial is appropriate may be judged by the F ratio. For this purpose the sum S_R^2 of squares of deviations of the experimental values y_i from the values $\hat{y_i}$ obtained by the regression equation (5.234)

$$S_R^2 = \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 \tag{5.247}$$

is calculated.

Dividing this sum by the degrees of freedom of the experiment, $n' = n - (k + 1)$ yields the variance s_R^2 characterizing dispersion of experimental points about the regression equation

$$s_R^2 = S_R^2/n' \tag{5.248}$$

In addition, the variance s_y^2 is calculated that describes the error of the experiment

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{5.249}$$

where $\overline{y}$ is the sample mean over all observations, and $n - 1$ is the number of degrees of freedom in the evaluation of s_y^2.

Now, the F ratio is

$$F = s_R^2/s_y^2 \tag{5.250}$$

and the hypothesis may be tested about the goodness of fit by the given polynomial of the experimental results by comparing the calculated value of F with that found in the table of F distribution for degrees of freedom n' and $n - 1$.

5.9.3 The Notion of Experimental Design

In order to carry out an active experiment, the designer should plan its structure to achieve the purpose of the experiment at least expenditures and in shortest possible time. This approach is collectively called *experimental design*. We illustrate the advantages it offers by way of a simple example [27].

Example 5.26. A spring balance is used to weigh three pieces A, B, and C. Denote by X the unknown weight of the balance dish.

A common approach is to use four weightings, for the dish, and for each piece in turn. It is represented in Table 5.4, where a plus

Table 5.4

	X	A	B	C	Result
Common method	+	−	−	−	y_0
	+	+	−	−	y_1
	+	−	+	−	y_2
	+	−	−	+	y_3
Modified method	+	−	−	+	y_1
	+	+	−	−	y_2
	+	−	+	−	y_3
	+	+	+	+	y_4

and minus imply the presence of the respective weight on the dynamometer. We shall refer to such a table as a *matrix of experimental design*. Each piece thus requires two treatments to be weighted, once with the dish and the other time for the dish alone. Hence the weight of piece A will be $P_A = y_1 - y_0$. Denote by σ^2 the variance of error in each weighting, then the variance of error in the evaluation of P_A will be $\sigma_A^2 = \sigma_{y_1}^2 + \sigma_{y_0}^2 = 2\sigma^2$.

Another, modified, method of weighting is represented by the second experiment matrix. Here, the first three experiments are weightings of each piece alone, and the last experiment weights all the three pieces together. Now, the weight of, say A, will be determined as

$$P_A = 0.5\,(y_2 + y_4 - y_1 - y_3)$$

with the error variance $\sigma_A^2 = 4\sigma^2/4 = \sigma^2$.

The variance of error in the second design is half as large as in the first, that is, the second design is preferable. This fact is explained

by the number of experiments involved in the evaluation of each weight — two in the first, and all four in the second method. Thus, the last method ensures a better compensation of errors in individual weightings.

The question to be answered is therefore how can we arrive at the most efficient design of the experiment under study? This question is answered in numerous treatises on experimental design (see, e.g. [28, 29]). Specifically, in regression analysis, the efficient methods are *full factorial design* and *partial replication.*

5.9.4 Full Factorial Design

Assume that we deal with two independent input variables (called factors) x_1 and x_2 of a controlled process or plant, each of which being varied at two levels, conventially denoted by $+1$ and -1, or simply by $+$ and $-$. By way of example, if these factors are temperature ($x_1 = 120°C$ and $80°C$) and pressure ($x_2 = 3$ and 2 MPa), then an experiment carried out at 120°C and 2 MPa will be denoted as $(+, -)$.

Suppose that we are interested in a regression equation that reflects the effect on the output (dependent) variable y of not only factors x_1 and x_2, but also their interaction reflected by the product x_1x_2, viz.,

$$\hat{y} = b_0x_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 \tag{5.251}$$

where as before $x_0 = 1$.

It is an easy matter to see that all possible combinations of the two factors x_1 and x_2 varied at two levels will be exhausted by four experiments illustrated in Table 5.5. This table also contains the

Table 5.5

x_0	Design		x_3 / x_1x_2	y	Code
	x_1	x_2			
+	−	−	+	y_1	(1)
+	+	−	−	y_2	a
+	−	+	−	y_3	b
+	+	+	+	y_4	ab

column for x_0 that is always $+1$ and a column, called block for the product x_1x_2. The meaning of the code designation column will be explained later. The values of x_1 and x_2 set in individual experiments represent collectively the design of this experiment.

This experimental pattern is called a *full factorial design* because it has used completely all possible combinations of levels for both factors. It is denoted by 2^2 for two levels of design for two factors. If the design is constructed at m levels for n factors, then the notation will be an m^n design. In our further consideration we limit ourselves to the case of two levels $m = 2$, i.e. to designs of the 2^n type.

Treatment level combinations of a factorial experiment are denoted by lower case letters that refer to the high level of a particular factor (denoted by a plus in the table). If all factors are at the low level, the symbol (1) is used (it should be noted that, in fact, these levels may be low and high on some scale or simply an arbitrary distinction). The four treatment combinations for our simple 2^2 experiment, assuming a and b represent treatment level combinations for factors x_1 and x_2, are as follows:

$$(1),\ a,\ b,\ ab$$

If we return to the weighting experiment, Table 5.4 is exactly of the same type as the matrix of the 2^2 experiment. Consequently, the efficient weighting resulted from using the design of the 2^2 experiment.

Let us consider the application of the results of the factorial experiments. In the above situation the number of treatment level combinations (four) is exactly the number of unknown regression coefficients, therefore the latter can be determined by the results of the full factorial experiment. This approach, however, would not leave any degree of freedom for testing of the hypothesis about the goodness of fit of the selected model (type of regression equation) to the actual process behavior.

If there are some grounds to believe that the effect of interaction of the factors, expressed by the term $b_{12}x_1x_2$ is insignificant, then we may let $b_{12} = 0$ and write the regression equation as

$$\hat{y} = b_0x_0 + b_1x_1 + b_2x_2 \qquad (5.252)$$

Now the four treatment level combinations will be used to determine the coefficients b_0, b_1, and b_2, and one degree of freedom remains for goodness-of-fit testing.

Let us consider the case of three factors A, B, and C, denoted by x_1, x_2, and x_3, that are varied at two levels, that is described by the 2^3 experiment. In order to test all possible treatment combinations we need to carry out eight tests. These combinations along with the *factor interactions* x_1x_2, x_2x_3, x_1x_3, and $x_1x_2x_3$ are summarized in Table 5.6. The symbolic designation for this experiment is

$$(1),\ a,\ b,\ ab,\ c,\ ac,\ bc,\ abc$$

Table 5.6

x_0	Design			x_4 $\widehat{x_1x_2}$	x_5 $\widehat{x_1x_3}$	x_6 $\widehat{x_2x_3}$	x_7 $\widehat{x_1x_2x_3}$	y	Symbol
	x_1	x_2	x_3						
+	−	−	−	+	+	+	−	y_1	(1)
+	+	−	−	−	−	+	+	y_2	a
+	−	+	−	−	+	−	+	y_3	b
+	+	+	−	+	−	−	−	y_4	ab
+	−	−	+	+	−	−	+	y_5	c
+	+	−	+	−	+	−	−	y_6	ac
+	−	+	+	−	−	+	−	y_7	bc
+	+	+	+	+	+	+	+	y_8	abc

where our method of designating these treatments is to include a particular lower case letter if the factor is at the high level and to exclude the letter if the factor is at the low level. Notice that the designation for the 2^3 experiment may be obtained from the designation of the 2^2 experiment repeated twice for the third factor C (x_3) at low (—) and high (+) levels respectively. If another factor (D) comes into play, then the treatment level combinations are repeated for three factors with the symbols for the fourth at the low and high levels.

The experiment carried out according to the design 2^3 yields the eight coefficients b_0, b_1, b_2, b_3, b_{12}, b_{13}, b_{23}, b_{123} but leaves no one degree of freedom for goodness-of-fit testing. In many situations, however, double and triple interaction effects are neglected so that the model is the linear regression

$$\hat{y} = b_0x_0 + b_1x_1 + b_2x_2 + b_3x_3 \tag{5.253}$$

This approach retains four degrees of freedom for goodness of fit testing.

To reveal other properties of the above designs we introduce new variable as indicated in Tables 5.5 and 5.6 to denote double and triple interactions. In this notation, the values of factor labelled i in level factor combination l possess the following properties

$$\sum_{l=1}^{n} x_i^{(l)} = 0, \quad i = 1, \ldots, k \tag{5.254}$$

$$\sum_{l=1}^{n} x_i^{(l)} x_j^{(l)} = 0, \quad i, j = 0, \ldots, k; \quad j \neq i \tag{5.255}$$

$$\sum_{l=1}^{n} (x_i^{(l)})^2 = n, \quad i = 0, \ldots, k \tag{5.256}$$

The last two equations imply that the matrix of experiment $\mathbf{X} = [x_i^{(l)}]$ defined by (5.239) is *orthogonal*, i.e. satisfies the condition

$$\mathbf{X}^T\mathbf{X} = n\mathbf{I}, \quad (\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{I}/n \tag{5.257}$$

In view of (5.246) the vector of regression coefficients in this case is

$$\mathbf{B} = \frac{1}{n}\mathbf{X}^T\mathbf{Y} \tag{5.258}$$

or

$$b_i = \frac{1}{n}\sum_{l=1}^{h} x_i^{(l)} y_l, \quad i = 0, \ldots, k \tag{5.259}$$

5.9.5 Fractional Replication

Let us consider in more detail the case where the effects of dual, triple, and higher interactions are neglected and a linear model is adopted. For the case of three variables, the linear approximation model is equation (5.253) involving four unknown regression coefficients whereas the full factorial experiment has eight level factor combinations. Four degrees of freedom are left for goodness-of-fit testing.

When a linear model is used, the number of unknown regression coefficients is $n = k + 1$ and the number of treatment combinations in a full factorial experiment $N = 2^k$ exceeds the number of unknowns and the difference rises steeply with k. For instance, for $k = 6$ we have $N = 64$, $n = 7$ and degrees of freedom $N - n = 57$. Clearly, in both situations the number of treatment combinations may be considerably reduced. The question to answer, however, is which of $N = 2^k$ experiments should be retained to keep the high efficiency of the whole experiment. The answer to this question is given by the idea of *fractional replications.*

Let us revisit the problem of three variables in which the linear effects are expressed by the four regression coefficients b_0, b_1, b_2, and b_3. Four treatment combinations are needed to determine them. What kind of experiments should they be?

Four experimental units can be obtained by carrying out a full factorial experiment for two variables only, say x_1 and x_2 in a 2^2 design. The levels of the third factor x_3 that should be given for each unit may be found by relating this factor with x_1 and x_2 in some way, say by letting $x_3 = x_1x_2$. The resulting design matrix for three factors completely coincides with that for the 2^2 experiment. It is described by Table 5.5, where the x_1x_2 column expressed the levels of the third factor. This table is seen to be one half of Table 5.6. This design is referred to as a 2^{3-1} experiment, for this experiment is planned for

three factors on the base of a complete factorial experiment for $3 - 1 = 2$ factors.

Each factor in Table 5.5 may be obtained as a product of two other factors.

$$x_1 = x_2x_3, \quad x_2 = x_1x_3, \quad x_3 = x_1x_2$$

This implies that we cannot tell the effect x_1 from the effect x_2x_3, that is the regression coefficient b_1 will be a common estimate for β_1 and β_{23}, written as $b_1 \to \beta_1 + \beta_{23}$. Likewise $b_2 \to \beta_2 + \beta_{13}$, and $b_3 \to \beta_3 + \beta_{12}$. No particular significance is attached to the estimates of β_{12}, β_{23}, and β_{13}, but if these quantities are other than zero, they will affect the accuracy of the estimate of β_1, β_2, and β_3.

The 2^{3-1} experiment may be planned in a different way, namely, by planning a complete factorial 2^2 experiment for two factors x_1 and x_2 and letting $x_3 = -x_1x_2$. In this design, the linear and paired effects, aliases, will be related as

$$x_1 = -x_2x_3, \quad x_2 = -x_1x_3, \quad x_3 = -x_1x_2$$

The regression coefficients b_1, b_2, and b_3 will be the estimates of the quantities $\beta_1 - \beta_{23}$, $\beta_2 - \beta_{13}$, and $\beta_3 - \beta_{12}$.

Two above methods of planning a 2^{3-1} experiment yield together a complete factorial experiment of the size of 2^3. Therefore these designs may be viewed as fractions of a full factorial experiment called *half-replicates*. Below we consider some principles fundamental to such replication.

In an experiment with five factors, the linear effects need six regression coefficients to be determined, and the number of experimental units in the full factorial experiment would be $2^5 = 32$. A half-replicate of this experiment would involve $2^{5-1} = 16$ observations, which is also a large number. Clearly, in this case the experiment may be limited to one fourth of the complete factorial design, denoted by 2^{5-2}. Smaller fractions, i.e. 2^{k-l} experiments may be derived in a similar manner.

Three principles basic to the concept of fractional replication are as follows.

1. Because the full factorial experiment is a most efficient type of experiment, a 2^{k-l} fractional experiment is based on a full factorial experiment with $k - l$ factors $x_1, \ldots, x_{k-l}$.

2. The remaining l factors $x_{k-l+1}, \ldots, x_k$ are expressed through the main factors in terms of confounded interactions. The choice of a *confounded interaction* completely defines the design of the experiment.

To plan a 2^{3-1} experiment we may employ one of the confounded interactions $x_3 = x_1x_2$ or $x_3 = -x_1x_2$.

In planning a 2^{4-1} experiment the full factorial experiment involves three factors x_1, x_2, and x_3. The number of confounded inter-

actions expressing x_4 through x_1, x_2, and x_3 is larger in this case; they are $x_4 = x_1x_2$, $x_4 = x_1x_3$, $x_4 = x_1x_2x_3$, $x_4 = -x_1x_2$, and so forth.

In a 2^{5-2} experiment, the confounded interactions relate x_4 and x_5 with x_1, x_2, and x_3 as

$$\begin{array}{ll} x_4 = x_1x_2 & x_5 = x_1x_3 \\ x_4 = x_1x_2 & x_5 = x_1x_2x_3 \\ x_4 = x_1x_3 & x_5 = x_2x_3, \text{ etc.} \end{array}$$

3. In choosing a preferable confounded interactions the experimenter should keep in mind that using a fractional replication approach he obtains not the regression coefficients but rather the estimates for some aliases (i.e. two effects or interactions that cannot be distinguished in analysis), as was shown above for a 2^{3-1} experiment. It is desirable that the design of an experiment should include only those aliases that affect the evaluated regression coefficients less than others. In order to arrive at a more appropriate design of an experiment, the experimenter ought to rapidly and easily evaluate the aliases for various designs. The concept of defining contrast is a useful tool to this end.

The *defining contrast* is a relation of factors that yields elements of the first column, i.e. $+1$. We illustrate the evaluation and application of a defining contrast by a 2^{3-1} experiment.

Denote the column-vector of the elements of the first column by $I = (1, \ldots, 1)^T$. As a confounded interaction we take $x_3 = x_1x_2$. Multiplying both parts by x_3 and observing that $x_3^2 = 1$, we obtain the defining contrast $I = x_1x_2x_3$. If we multiply this relation in turn by x_1, x_2, and x_3, we obtain the *alias pairs* $x_1 = x_2x_3$, $x_2 = x_1x_3$, and $x_3 = x_1x_2$ which were derived earlier by another argument.

Another example is a 2^{5-2} experiment. The main factors are x_1, x_2, and x_3. As confounded interactions we take $x_4 = x_1x_2$ and $x_5 = x_1x_2x_3$. Multiplying these relations by x_4 and x_5, we arrive at the defining contrasts $I = x_1x_2x_4$ and $I = x_1x_2x_3x_5$. If we multiply these defining contrasts pairwise, then we obtain a third relation $I = x_3x_4x_5$ with the elements of the first column.

In order to define all aliases we introduce the *generalized interaction*

$$I = x_1x_2x_4 = x_1x_2x_3x_5 = x_3x_4x_5$$

To obtain the aliases in determining the regression coefficient b_1 we multiply by x_1 all the terms of this expression

$$x_1 = x_2x_4 = x_2x_3x_5 = x_1x_3x_4x_5$$

which implies

$$b_1 \to \beta_1 + \beta_{24} + \beta_{235} + \beta_{1345}$$

Commonly the three and four factor interactions are fairly small, therefore we may deem $b_1 \to \beta_1 + \beta_{24}$.

Using this argument we can find aliases in determining other regression coefficients.

EXERCISES

5.1. What are events S_1 and S_{15} in Example 5.1?

5.2. Prove that if a sample description space consists of n elements, then it contains 2^n distinct events.

5.3. A warehouse receives fuses of identical rating from two production lines in the ratio 1 : 3. The first line is known to produce 10% and the second 20% of out-of-tolerance items. Use the theorem of total probability to obtain the unconditional probability that a fuse taken at random will be substandard.

5.4. Plot the graph of the probability distribution function for the number of points coming up in throwing two fair dice of Example 5.1.

5.5. Determine the probability of the waiting time in Example 5.8 being within 3 minutes.

5.6. In order to learn the population fish in a lake 1000 of netted fish is labeled and returned to water. What should be the number of fish in the lake for the probability of finding 10 labeled species in a sample of 150 be highest?

6

OPTIMAL CONTROL PROBLEMS

6.1 Basic Features of a Control Process

6.1.1 The Notion of Control

Processes that take place in the nature, technology, and society depend on a multitude of concomitant conditions and factors. By varying the conditions of a process one can influence its behavior and slave it to certain objectives. This interference into the natural development of a process with the purpose to alter its behavior is the essence of control. Saying it in other words, a *control* is an arrangement or management of a process, which ensures the attainment of certain objectives [30, 36].

For the purpose of vivid perception of the process of control let us consider an example of a dog hunting a hare. In order to overtake the hare, the dog should organize its actions in a certain way and control them. Consequently, the process of hunting is an example of control.

Before the hunting begins, the hare must appear in the dog's sight, i.e. there must be a situation with a certain object whose attainment is needed or desirable. Prior to begin the hunt, the dog must estimate the situation, compare it with its desires and capabilities. This estimation ends up in making a decision on whether or not it is worth hunting (the hare may be too far a target, or the dog may be tired by a previous hunting, etc.). When the decision to start hunting is made, the dog begins arranging its motion with the purpose to overtake the hare in shortest time or at least effort.

Four stages are evident in this example, typical of any process of control, namely, the appearance of a target, estimation of the situation, decision making, and realization of this decision. Since the appearance of the target or objective precedes the control process, we exclude it from the consideration. Recognizing also that in complex-process control the estimation of a situation is carried out on the basis of a gathered and appropriately processed information, we arrive at the three stages of control: (1) collection and processing of data to estimate the situation at hand; (2) making a decision on more appropriate actions; and (3) realization of the decision. In some situations a fourth stage may be necessary — verification of

the implementation. Various types of control problems differ from one another by the method and order in which this operations are effected.

6.1.2 Types of Control Problems

There are many applications in which the mechanisms of data collection and realization of a decision are perfected to an extent relieving the analyst from attention to these stages so that the control process reduces essentially to considering the second stage only. Such types of problem are referred to as *one-stage* or *single-step* decision making.

In many situations, however, this approach is an excessive simplification of the real process. In actual practice all the stages of control processes are in intimate relationship and the step of decision making requires more or less detailed consideration of the possible methods in which the decision being made may be implemented. For example, in order to make a decision that the hunting of the hare is not worth continuing, the dog must convince itself that the hunt is of no use. To do this, it must at least roughly analyze all possible methods of hunting.

Sometimes, in similar cases the process of control is broken down in a few successful steps so that the decision made at some step depends on the results of the decision at the preceding step. Such processes are termed the *processes with sequential decision making.* An example may be the process of guiding a rocket launched from earth to moon. Steps to be singled out are putting the rocket into an earth's orbit, guidance of the rocket in its motion to the moon, control of the rocket into a lunar orbit, and landing on the moon.

In this example steps of control are self-evident. However, in many a situation, partitioning of a complex process of control into steps with clearcut indication of all the stages of control at each step is a tall task. For example, in hunting a hare we deal with a continuously changing situation due to the hare wishing to go away from the hunter. The dog should be assessing the situation and take ever new decisions according to the situation, without having to wait for the results of the previous decisions. This is a *continuous dynamic process of control.*

The situations considered above are indicative of how complicated and diverse control problems may be. We would, however, underestimate the difficulties involved in solving these problems if we fail to observe that the processes of control are as a rule embedded in a complicated environment. They undergo the influence of diverse external factors, which are often collectively referred to as the *state of the nature.* In order to make a correct decision about some actions we should estimate the results due to these decisions, and for this

we should know the situation in which these actions are to be performed.

However, a typical situation for control problems is one with the available information either insufficient for an accurate assessment of the situation, or distorted by extraneous factors. Uncertainty by no means makes the problem of control easier and the decision has to be made irrespective of whether or not we are able to estimate correctly the results of this decision. This problem is called *decision making under uncertainty.*

6.1.3 The Notion of Operations Research

One more specific class of control problems is due to the management of industrial enterprises and large ventures [31, 32].

Prior to the industrial revolution a small enterprise could have been managed by a single manager who did all buying, planning of jobs, selling and hiring. He could make decisions without having to consult any scientific methods — just basing on his own experience and intuition. If some of his decisions were not best, they could not do much harm, or could rapidly be corrected.

Large enterprises of today can no longer be managed by one person. The functions of buyers, sellers, production and personnel are now divided among several managers. Moreover, the management of production department in large works may be divided into still smaller groups dealing with production, tool repair, quality control, planning and scheduling, supply, storage, and the like.

Each specialized department of a large enterprise performs a proportion of a common task as required by the general objective of the enterprise. In doing so it aims at its own objects which do not always agree with, and sometimes contradict other objectives.

As an example we consider an inventory problem. Some department or shop may be interested in considerable growth of stock on the shelves in order to secure continuous production. However, under a limited space in the storehouse this desire may lead to the reduction of stock in store for other departments or shops. The situation requires that a strategy should be developed concerning the stock in store such that would be favorable for the entire enterprise.

In solving this type of managerial problems the analyst needs to evaluate the objectives of individual departments or shops and see that they do not contradict each other and the general objective of the enterprise. Recognizing that a nonoptimal decision may do harm larger the larger the enterprise, it is clear that modern production planning should not be based on personal experience and common sense. The appropriate mathematical methods developed to assist the manager involved in production planning and similar problems are collectively termed *operations* or *operational research.*

The *operation* is understood as an endeavor aimed at a certain objective, for example, at a certain amount and proportion of stock in the warehouse. The conditions describing the limitations of the endeavor, specifically the stock requirement and constraints on storage space, should be specified. The purpose of operations research is to evaluate such methods of carrying out the endeavor that would be most favorable in a certain sense.

Production planning problems are specific in that the effect of one or another method of their solution may have severe consequences for the performance of the entire enterprise. Therefore, a final decision is made by a manager, referred to as a decision maker or DM. He has a power to decide which solution of those evaluated by the analyst is best for the specific situation under certain circumstances. Operations research is aimed therefore at providing the DM with well based recommendations on decisions and their consequences. On the way to these recommendations the analyst develops the methods of analysis of targetted actions (operations) and objective comparison of feasible solutions. Essentially operations research consists of the same mathematical methods that constitute what is called *cybernetics* in this volume.

6.2 Optimization of Control Processes

6.2.1 Performance Measure

In what follows we shall treat the problem of control as a mathematical problem. Unlike many other mathematical problems this one allows not single but many different solutions [33]. This feature stems from the fact that in control problems there are as a rule many ways in which some process may be guided to a certain objective. For example, in persuing a hare the dog may use different pace and directions, i.e. use different patterns of motion; in launching a rocket to the moon the ground control can choose different trajectories, and such. Therefore, a problem of control may be formulated as a search for a feasible path of attaining to the set objective. Usually, however, this formulation proves to be insufficient.

If there is a set of solutions to some optimization problem, then we should select the one which is best from some standpoint. For example, there are many ways to make a box of a sheet of cardboard; the solution of this problem is a box of the largest capacity. From one city to another a passenger can travel by air, water, bus, and car. The optimal solution to this problem is a transportation means that is most favorable from the standpoint of the time of travel, cost of fare, comfort, etc. Similar situation takes place in control problems.

When the objective can be attained by several different ways, additional constraints may be imposed on the method of control so that the degree in which these requirements are met may be used as a yardstick to choose a method of control.

In many situations, realization of a control process requires the expenditure of certain resources, such as time, materials, fuel, or electrical energy. Accordingly, in choosing a method of control we should speak not only of whether or not the objective is achieved, but also of what resources will have to be expended to attain this objective. In this case the control problem consists in selecting, from a set of feasible solutions, i.e. those leading to the objective, the one which requires less resources than others.

In other situations the basis upon which a method of control is selected may be formed by other constraints imposed on the control system, such as maintenance cost, reliability, closeness of the system state to the required, the degree of significance in the knowledge of the state of nature, and the like.

A mathematic expression yielding a quantitative estimate of how the constraints imposed on the method of control are met is called the *objective function*, *figure of merit*, or *performance measure* (*criterion*). The most favorable or optimal method of control is the one that maximizes (minimizes) the objective function.

By way of illustration, in choosing a flight regime for a craft one may take as a performance measure the expression for fuel expended per unit path, or a length traveled by firing a unit amount of fuel. The most economical, i.e. optimal, regime will be the one corresponding to an extremal value of the objective function.

We shall consider this definition of optimal control as a preliminary one. A more rigorous definition will be given after a consideration of the constraints imposed on the control process.

6.2.2 Constraints

A formulation of the problem of optimal control or control in general should be deemed inexistent if no constraints are imposed on the behavior of the system in question. For example, there were no problem of hare hunt if the dog could cover the distance to the hare in as short time as desired. Consequently, in dealing with a control problem we should not overlook that the behavior of any system is subject to various constraints.

To provide a better insight into the types of constraints frequently encountered in control problems we consider a specific example of driving. The driver should keep in mind that the engine has a finite power, consequently under a certain load it can deliver to the truck a certain maximum speed. Owing to the inertia of the vehicle its speed and direction of motion can be changed with a certain acce-

leration. This implies that it cannot be brought to a halt in a very short time or suddenly change its direction of travel when such a need occurs. In careful riding these limitations dictate a maximum speed. When selecting a route the driver should observe the capacity of fuel tank and the location of gas and oil stations along the route.

In general there are two types of constraints imposed on the method of control [30]. The first type of constraint occurs due to the laws of nature which govern the behavior of the controlled system. In mathematical formulations of control problems these constraints are usually represented by algebraic, differential or difference constraint equations. The second type of constraints is due to the limited amount of resources used in the process or due to some other limited quantities that owing to the physical nature of the system should not exceed certain limiting values. Mathematically this type of constraint is represented by systems of algebraic equations or inequalities that relate the variables describing the states of the system.

6.2.3 Problem Statement of Optimal Control

A problem of control may be deemed mathematically formulated if (1) the purpose of control is formulated in terms of a performance measure, (2) the constraints of the first kind are determined as systems of differential or difference equations describing all feasible methods of system motion, and (3) the constraints of the second type are delineated as algebraic equations or inequalities representing the limited amount of resources or other quantities involved in the control process.

A method of control that satisfies all constraints imposed on the process and minimizes (maximizes) the performance measure of the process is termed *optimal control.*

6.3 Mathematical Description of the Controlled System

6.3.1 Production Planning Models

Mathematical models for analysis and planning of production activity of individual enterprises, industries, or economic systems are often referred to as *production planning models.*

The controlled system, process or plant may be considered as a many port system whose inputs may be categorized as follows: **X**, technological inputs that represent raw materials, stocked items, semifinished work pieces, etc.; **R**, tools, machines and other equipment necessary for processing; and **L**, labor involved in the production process.

The output **V** of the production process is the amount of finished product.

We shall call the dependence of the output upon the inputs the *production function* written generally as

$$\mathbf{V} = F(\mathbf{X}, \mathbf{L}, \mathbf{R}) \tag{6.1}$$

The type of this dependence is very diversified and depends upon specific objectives of the investigation. We shall confine ourselves to considering the simplest type of this dependence — a *linear production function*, in which the model output is a linear function of the inputs. It should be kept in mind that in any real production process each group of inputs as well as the output will be multicomponent, i.e. vectorial, quantities, viz.,

$$\begin{aligned}\mathbf{X} &= (x_1, \ldots, x_n)^{\mathrm{T}}\\ \mathbf{R} &= (r_1, \ldots, r_k)^{\mathrm{T}}\\ \mathbf{L} &= (l_1, \ldots, l_m)^{\mathrm{T}}\\ \mathbf{V} &= (v_1, \ldots, v_r)^{\mathrm{T}}\end{aligned}$$

We introduce the coefficients

$$\begin{aligned}a_{ij} &= x_i/v_j\\ b_{ij} &= l_i/v_j\\ h_{ij} &= r_i/v_j\end{aligned} \tag{6.2}$$

that define the amount of resource i of **X**, **L**, or **R** expended to produce one unit of product j. These coefficients can be conveniently arrayed into the matrices $\mathbf{A} = [a_{ij}]$ of material resources, $\mathbf{B} = [b_{ij}]$ of labour resources, and $\mathbf{H} = [h_{ij}]$ of production facilities.

Now, the linear objective function can be written as matrix relationships

$$\mathbf{X} = \mathbf{AV}, \quad \mathbf{L} = \mathbf{BV}, \quad \mathbf{R} = \mathbf{HV} \tag{6.3}$$

This model, however, fails to take into account the final objective of the production process, which in this case is to satisfy the demand of the consumer market. The objective of production planning is normally to strike a certain balance between the production and consumption. This balance is expressed in the form of balance equations describing multiple product production models. These models may be static, where the process is assumed invariable over a rather long time, or they may be dynamic, that is, assuming continuous growth and improvement of the production process. We shall briefly touch upon both models.

Static multiproduct model. Consider the case of n enterprises producing n types of product, denoted by $v_1, \ldots, v_n$. Each type of product v_i is consumed at a rate w_i on the market and is expended at a rate x_i in the production of other types of product. This fact may be reflected by the balance equation

$$v_i = w_i + x_i, \quad i = 1, \ldots, n \tag{6.4}$$

Under a linear model, the consumption x_{ij} of product i in process j is proportional to the output of product j, so that

$$x_i = \sum_{j=1}^{n} x_{ij} = \sum_{j=1}^{n} a_{ij} v_j, \quad i = 1, \ldots, n$$

The set of all balance equations for n production processes is

$$v_i - \sum_{j=1}^{n} \alpha_{ij} v_j = w_j, \quad i = 1, \ldots, n \tag{6.5}$$

These equations may be written in a compact form by introducing a consumption vector $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$ and the coefficient matrix

$$\mathbf{C} = \begin{pmatrix} 1 - a_{11} & -a_{12} & \ldots & -a_{1n} \\ -a_{21} & 1 - a_{22} & \ldots & -a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ -a_{n1} & -a_{n2} & & 1 - a_{nn} \end{pmatrix} \tag{6.6}$$

Now the system of equations (6.5) will be written as one matrix equation

$$\mathbf{CV} = \mathbf{W} \tag{6.7}$$

Solving this system of equations for $\mathbf{V}$, we obtain the production rates satisfying the specified consumption

$$\mathbf{V} = \mathbf{C}^{-1}\mathbf{W} \tag{6.8}$$

or, denoting by α_{ij} the elements of the inverse matrix, $\mathbf{C}^{-1}$,

$$v_i = \sum_{j=1}^{n} \alpha_{ij} w_j, \quad i = 1, \ldots, n \tag{6.9}$$

In order to ensure the production output with intensities defined by (6.9) we need the resources

$$\begin{aligned} L &= \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} b_i v_i \\ R &= \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} h_i v_i \end{aligned} \tag{6.10}$$

Because the final objective of the production is to satisfy the market demand, the quantities L and R would be convenient to express in terms of consumption rate w_j, rather than production ra-

tes v_i. This purpose can be achieved with the reduced coefficients of labor and material resources

$$\bar{b}_j = L_j/w_j \text{ and } \bar{h}_j = R_j/w_j \tag{6.11}$$

The connection between the coefficients b_i, h_i and $\bar{b}_j$, $\bar{h}_j$ may be readily established by substituting v_i from (6.9) into (6.10), namely,

$$L = \sum_{i=1}^{n} b_i \sum_{j=1}^{n} \alpha_{ij} w_j = \sum_{j=1}^{n} \bar{b}_j w_j \tag{6.12}$$

where

$$\bar{b}_j = \sum_{i=1}^{n} \alpha_{ij} b_i \tag{6.13}$$

Similarly,

$$R = \sum_{j=1}^{n} \bar{h}_j w_j \text{ with } \bar{h}_j = \sum_{i=1}^{n} \alpha_{ij} h_i \tag{6.14}$$

In actual situations the relationships (6.12) and (6.14) cannot always be satisfied, because labor and other resources, such as production equipment may be limited. If the amount of labor is limited by the number L_0 and the resources of production equipment by R_0, then both (6.12) and (6.14) should be replaced with the inequalities

$$\sum_{i=1}^{n} \bar{b}_i w_i \leqslant L_0 \text{ and } \sum_{i=1}^{n} \bar{h}_i w_i \leqslant R_0 \tag{6.15}$$

where v_i and w_i are subject to the constraints that these quantities be nonnegative

$$v_i \geqslant 0, \quad w_i \geqslant 0, \quad i = 1, \ldots, n \tag{6.16}$$

We shall also assume that the products of our enterprises suit the same purpose, and denote by p_i the cost or utility of product i. In this setting the production planning problem may be formulated with the objective to find such production rates that maximize the total worth of the product

$$f(w_1, \ldots, w_n) = \sum_{i=1}^{n} p_i w_i \tag{6.17}$$

This problem is solved by the linear programming methods and therefore is called an LP Problem.

Dynamic multiproduct model. Static models do not reflect the fact that the production process is developing and improving. Another class of models that reflect this circumstance are referred to as *dynamic models.*

Capital investments represent the internal forces that provide the development of the production process. These investments can be

obtained at the expense of the product, V, being produced, and accumulated and form an *accumulation fund*. The other part of the product forms a *consumption fund*.

It is conditional to part the accumulation fund in two. The first part H_R makes up the production fund expended for the improvement of the production means. The second part H_i is intended for the enhancement of the information potential of the production that includes the capital investments for research and development and for education of the personnel.

The effect of capital investment always takes time to be obtained. Capital investments for the expansion of production funds H_R give an effect as a rule with a shorter delay, however, this payoff is smaller. Expenditures for the development of science and personnel education are realized with large delay, however they may dramatically change the character of the production process and take care of the consistent improvement of the technology.

In what follows consider for simplicity one accumulation fund and assume that the effect of capital investment realizes immediately. We denote by $g_i(t)$ the rate of product going to the accumulation fund in the i-th production process. The balance equation for enterprise i will be

$$v_i(t) = w_i(t) + x_i(t) + g_i(t) \tag{6.18}$$

This equation indicates that the product v_i is consumed at a rate w_i, used for reproduction at a rate x_i and expended for the growth of the production funds at a rate g_i. If we denote by x_{ij} the rate of expenditure of product i for the reproduction of product j, and by g_{ij} the rate of expenditure of product i for the capital investment in the production of product j, then we obtain

$$x_i = \sum_{j=1}^{n} x_{ij} = \sum_{j=1}^{n} a_{ij} v_j, \quad g_i = \sum_{j=1}^{n} g_{ij} \tag{6.19}$$

In order to combine the expenditure of a product for the increase of the production fund with the growth of the production we need to merge two processes — the rise of the production fund H_{ij} and its expenditure. Consider the increment of the production fund $dH_{ij}(t)$ achieved over a small interval dt. This increment is proportional to the accumulation rate $g_{ij}(t)$ and the interval dt.

$$dH_{ij}(t) = c_{ij} g_{ij}(t)\, dt \tag{6.20}$$

The expenditure of the production fund goes for the improvement of the tools used $R_{ij}(t) = h_{ij} v_j$, therefore

$$dH_{ij}(t) = dR_{ij}(t) = h_{ij} dv_j(t) \tag{6.21}$$

Comparing (6.21) and (6.20), we find $c_{ij}g_{ij}(t)\,dt = h_{ij}dv_j(t)$. Whence

$$g_{ij}(t) = k_{ij}dv_j(t)/dt \tag{6.22}$$

where $k_{ij} = h_{ij}/c_{ij}$ is the relative capital investment.

In view of (6.19) and (6.22) the balance equation (6.18) rewrites as

$$v_i(t) - \sum_{j=1}^{n} a_{ij}v_j(t) - \sum_{j=1}^{n} k_{ij}\,dv_j(t)/dt = w_i(t),\ i = 1,\ \ldots,\ n \tag{6.23}$$

This problem is extremely hard to solve in view of a large number of unknowns involved. It is aggravated by the need to observe the constraints on labor and tools (6.15) and the conditions (6.16) of the variables being nonnegative. We shall not pose on the methods of solution of this system of equations to be found elsewhere [35].

6.3.2 Structure of the Controlled System

The production models discussed in the previous subsection are helpful in solving many problems of production planning. Planning, however, is but one stage of a control process. The other stage, realization of the plan, involves the processes of control of technological and production facilities. To solve this problem one needs a more detailed description of specific controlled systems or processes (plants) in the form of mathematical models.

We call the physical system whose processes are subject to control the *plant* or *process to be controlled.* Plants may be of various physical nature such as a car, airplane, rocket, machine tool, etc. or these may be production facilities such as a department, shop, works, or industry, or these may be economic systems when one speaks of managing the process or industry economy or even the economics of the state; finally these may be biological systems, social systems and such.

The regularities which the plants and processes obey are common for controlled objects of any physical nature. Therefore we are in a position to provide a common mathematical formalism of the control process.

Denote by x a variable that defines the state of the plant. Sometime it is one-dimensional or scalar. This variable may be the angle of rotation of a motor shaft, the velocity of an airplane or a rocket, the pressure of steam in a steam engine boiler, the number of items in a warehouse, the number of aircraft in an airfield, etc.

Most often than not, however, a few variables rather than one are required to describe the plant. We denote them by $x^{(1)}, \ldots, x^{(n)}$. In dealing with mechanical systems the quantities $x^{(i)}$ represent the coordinates or velocities of the moving parts, in electrical systems

they are current or voltage, in economics these are the capacities or resources of the particular industries, and in a biological system these quantities may characterize the concentration of chemical substances or medicals in different tissues.

Under the circumstances the state of the object under discussion will be described by a multidimensional variables $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})$ which will be treated further as a point in space R^n. The variable x will be referred to as the state *variable* or the *state vector* of the controlled process or plant. As a rule the quantities $x^{(i)}$ and hence the vector x cannot take on any values. The set of feasible values of x will be denoted by X and will be referred to as a *feasible set* or *range of feasible solutions*. In doing so we emphasize that the choice of some x in X is a feasible solution to the control problem.

If the quantities $x^{(i)}$ assume a finite set of values, then the feasible set X will be also finite, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with $\mathbf{x}_k = (x_k^{(1)}, \ldots, x_k^{(n)})^{\mathrm{T}}$ and $k = 1, \ldots, N$.

If the quantities are allowed to vary continuously, i.e. may assume an infinite set of values, then the feasible set X will be infinite. However, in this case as well the values of $x^{(i)}$ are not arbitrary. They are subject to some constraints that, as we have already mentioned, have the form of algebraic equations or inequalities, viz.,

$$f_i(x^{(1)}, \ldots, x^{(n)}) \begin{cases} \leqslant b_i, \text{ or} \\ = b_i, \\ \geqslant b_i \end{cases} \tag{6.24}$$

with $i = 1, \ldots, m$.

Each of the constraints (6.24) retains only one sign, $\leqslant$, $=$, or $\geqslant$, however various constraints may have different signs.

The quantities m and n are not related to one another, so that m may be more than, equal to, or less than n. Specifically, m may be zero, therefore the case when (6.24) are absent is not excluded. More often than not some or all the variables $x^{(i)}$ satisfy the nonnegativity condition $x^{(i)} \geqslant 0$.

The condition that the variables be nonnegative is a convenient tool in handling the equations of control by numerical methods. In addition, many problems, say economical, physically preclude $x^{(i)}$ to be negative. Indeed, such quantities as expenditures, production capacities, amounts of shipments, and allocated sums of money cannot be negative. However, when the variables can be of arbitrary sign, i.e. obey the constraints

$$x^{(i)} \geqslant a_i, \quad i = 1, \ldots, n$$

where a_i are arbitrary numbers, they can be readily transformed into nonnegative variables by the substitution $y^{(i)} = x^{(i)} - a_i$.

The state of controlled process depends not only upon variable x,

which will be assumed to be measurable and subject to controlled variations, but also upon a number of uncontrollable or not completely controllable factors governed by a collection of external conditions in which the process is embedded. These factors may be the composition of raw material, random perturbations of the environment, vibrations, and the like. In engineering such factors are commonly called *disturbances.*

The pilot can control the flight, for example, by varying the altitude and speed of flight, which in this case are controlled variables. On the other hand, the consumption of fuel depends considerably on the atmospheric conditions of the flight, which the pilot can take into account but which he cannot alter or even accurately forecast.

The whole set of uncontrollable external factors that influence the control process will be referred to as the *state of nature* and denoted by ϑ. In many a situation, with some simplification of real phenomena the uncountable multitude of external conditions may be reduced to a finite number of possible states of nature $\vartheta_1, \ldots, \vartheta_L$, i.e. to consider a finite set $\Theta = \{\vartheta_1, \ldots, \vartheta_L\}$ that will be called the *space of states of nature.* The members ϑ_i of the set Θ are in general multidimensional variates $\vartheta_i = (\vartheta_i^{(1)}, \ldots, \vartheta_i^{(l)})^{\mathrm{T}}$, $i = 1, \ldots, L$.

Since not all external factors are under complete control, in place of exact knowledge of the state of nature ϑ we have to restrict ourselves in many situations to merely the knowledge of the probabilities $\xi(\vartheta)$ of various states of nature $\vartheta \in \Theta$. The probabilities $\xi(\vartheta)$ that have been obtained by some way or another for all $\vartheta \in \Theta$ prior to problem solving will be referred to as the *prior probabilities* of the states of nature. It is quite natural that prior probabilities $\xi(\vartheta)$ representing the distribution of probabilities over space Θ must satisfy the conditions (5.3).

In some situations the state of nature ϑ should be considered a a continuous random variable for which the space of states of nature Θ is an infinite set. In this cases the distribution of probabilities $\xi(\vartheta)$ becomes a probability density function.

To describe the action that slaves the controlled system we introduce a variable u that will be referred to as controlled action or simply control. Thus, the term control will be further used to mean (1) control as management activity targetted to attainment of certain objectives, and (2) a control in the sense of control action, i.e. a physical quantity that is varied at our will and alters the processes in the control system in the desired way. The specific sense in which the term control will be used in a specific situation will be clear from the context.

Complex systems have to be controlled with a few control actions $u^{(1)}, \ldots, u^{(r)}$ so that the control in the general case is a multidimensional variable $\mathbf{u} = (u^{(1)}, \ldots, u^{(r)})^{\mathrm{T}}$. By way of example, the

pilot can alter the mode of flight by using a control u consisting of a fuel flow control (throttle) $u^{(1)}$, and an elevator control $u^{(2)}$.

In actual systems control actions $u^{(i)}$ are subject to various constraints. Denote by U the set of all values of control that satisfy the constraints, so that any u in U is a *feasible control*. In what follows we shall often meet problems in which U is a finite set, namely,

$$U = \{\mathbf{u}_1, \ldots, \mathbf{u}_R\} \tag{6.25}$$

How efficient a particular control u is in attaining the set objectives can be judged by the values of state variables x. However, in many cases this way is inconvenient because the objectives of control may be rather complex functions of the variables of state. It will be convenient therefore to introduce the *output variables* q that express the objectives of control in an explicit form. Sometimes a state variable may be used as an output variable, but this is not the case in general. For instance, to minimize the fuel consumption in a flight through a given route it is convenient to take as output variable the path covered with a unit amount of fuel; then the state variables will be the speed and altitude of the aircraft. Economical figures of merit can be also used as output variables, for example efficiency, or control performance index.

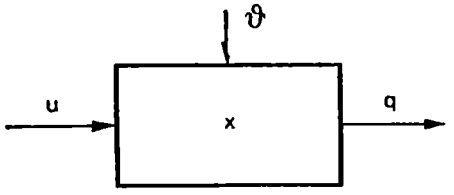


Fig. 6.1. Block diagram of a controlled system

The output variable q depends primarily upon the state of the controlled system, x. However, it may also come under the influence of control u and the external uncontrollable factors v. Hence, in the general case the equation for the output variable has the form

$$q = q(\mathbf{x}, \mathbf{u}, \vartheta) \tag{6.26}$$

The interaction of all variables considered above may be schematically represented by the graph of Fig. 6.1.

6.3.3 Equations of Motion

The signals of control u compel the controlled system to change its state. The relevant processes are defined by the variation rate of state variable $\dot{\mathbf{x}} = d\mathbf{x}/dt$ that is multidimensional

$$\dot{\mathbf{x}} = (\dot{x}^{(1)}, \ldots, \dot{x}^{(n)})^{\mathrm{T}} \tag{6.27}$$

where $\dot{x}^{(i)}$, $i = 1, \ldots, n$, are the variation rates of the components of x.

For dynamic systems where the physical processes are continuous in time, the rates $\dot{x}^{(i)}$ at some time instant depend upon the state of the controlled system at the same time instant, which in turn is defined by the values of the state variable x, the state of nature ϑ, and the control u used. This dependence may be written in the form of a system of ordinary differential equations

$$\dot{x}^{(i)} = g_i(\mathbf{x}, \mathbf{u}, \vartheta), \quad \mathbf{x}^{(i)}(0) = c_i, \quad i = 1, \ldots, n \tag{6.28}$$

where c_i, $i = 1, \ldots, n$, characterize the initial state of the controlled system.

Sometimes it is convenient to employ vector notation in which the system (6.28) becomes a single equation

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}, \mathbf{u}, \vartheta), \quad \mathbf{x}(0) = \mathbf{c} \tag{6.29}$$

where $\mathbf{c} = (c_1, \ldots, c_n)^{\mathrm{T}}$, and $\mathbf{g} = (g_1, \ldots, g_n)^{\mathrm{T}}$ is a vector function.

As a rule the output variables q are connected with the state variables by algebraic equations

$$q = f(\mathbf{x}) \tag{6.30}$$

If $\dot{\mathbf{x}} = 0$, the system model becomes

$$\mathbf{g}(\mathbf{x}, \mathbf{u}, \vartheta) = 0 \tag{6.31}$$

The presence of ϑ as an argument in (6.29) is inconvenient in a number of cases because the state of nature often remains invariable over the time of one control process. However, for different control processes the state of nature may be different. It manifests itself in different types of the equation of motion. This dependence of the type of equation on the state of nature will be denoted by the subscript ϑ at g so that (6.29) rewrites as

$$\dot{\mathbf{x}} = \mathbf{g}_\vartheta(\mathbf{u}, \mathbf{x}), \quad \mathbf{x}(0) = \mathbf{c} \tag{6.32}$$

If the state of nature does not change over the time of all control processes under consideration, then this index will be omitted.

6.3.4 Linear Models of Dynamic Systems

Because the variables $\mathbf{x}(t)$, $\mathbf{u}(t)$, and $\vartheta(t)$ are multidimensional, i.e. vectorial quantities, equation (6.28) is also vectorial and hence a rather complex one. The situation becomes even more involved if g is a nonlinear function. The theory of solution of such vector differential equations has been well developed for linear systems only,

therefore we limit our consideration to the linear dynamic models of controlled systems assuming that a nonlinear g can be linearized.

For a linear function g equation (6.28) may be represented in the vector form

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\,\mathbf{x}(t) + \mathbf{G}(t)\,\mathbf{u}(t) + \mathbf{\Gamma}(t)\,\vartheta(t) \qquad (6.33)$$

Here $\mathbf{F}(t)$, $\mathbf{G}(t)$, and $\mathbf{\Gamma}(t)$ are the transformation matrices whose elements may in general be functions of time. If the controlled sys-

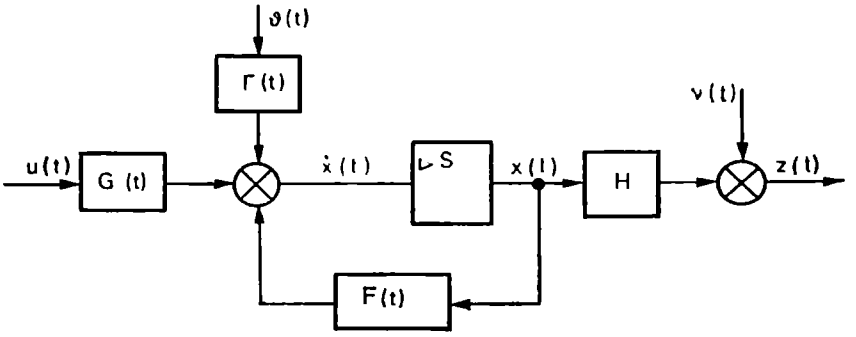


Fig. 6.2. Block diagram of a linear dynamic system

tem is time invariant, then the matrices **F**, **G**, and **Γ** will be constant.

To follow the behavior of the controlled system we install in it transducers that provide numerical estimates of all or some of the state variables. The output signals may be described by the matrix equation

$$\mathbf{z}(t) = \mathbf{H}\mathbf{x}(t) + \boldsymbol{\nu}(t) \qquad (6.34)$$

where $\mathbf{z}(t)$ is the measurement vector, **H** is the matrix relating the vector of state to the measurement vector, and $\boldsymbol{\nu}(t)$ is the vector of measurement errors, that is often referred to as *disturbance*.

The block diagram of a continuous linear system is shown in Fig. 6.2.

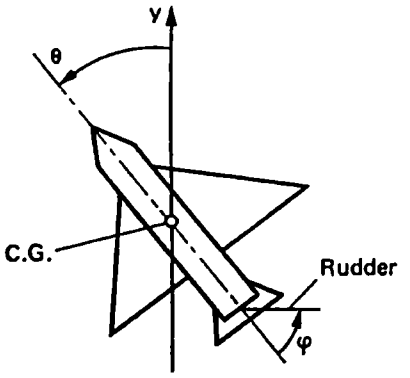


Fig. 6.3. Control by yawing angle

Example 6.1. Control of an airplane by yawing angle. Assume that gusts of wind have deflected the airplane axis by an angle θ from the prescribed direction y as shown in Fig. 6.3. The plane can be returned to the given course by deflecting the rudder through an angle φ. We assume that the moment of inertia about the center of

gravity of the plane is J, and the restoring force is proportional to φ, the viscous friction of air being neglected.

By the second Newton's law the equation of motion is

$$J\ddot{\theta} = -k\varphi(t) + m(t)$$

where $k\varphi(t)$ is the restoring force, and $m(t)$ the moment induced by the gusts of wind. Dividing this equation by J, denoting $b = -k/J$, and $\vartheta(t) = m(t)/J$, and observing that $\varphi(t)$ is the control action $u(t)$ yields

$$\ddot{\theta} = bu(t) + \vartheta(t)$$

On introducing the state variables $x_1 = \theta$ and $x_2 = \dot{\theta} = \dot{x}_1$, we arrive at two differential equations of the first order

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = bu(t) + \vartheta(t)$$

which may be combined into the vector form

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix} u(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \vartheta(t)$$

or

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x}(t) + \mathbf{G}u(t) + \mathbf{\Gamma}\vartheta(t)$$

where the matrix notation is self-evident.

If we assume that there are measuring instruments aboard that keep tracking the deviation in yawing angle $\theta = x_1$ and its rate $\dot{\theta} = x_2$, and denote the indications of these devices by z_1 and z_2 and the measurement errors by ξ_1 and ξ_2, then the equations of measurements are

$$z_1 = x_1 + \xi_1$$
$$z_2 = x_2 + \xi_2$$

which combine in one vector equation (6.34)

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \boldsymbol{\nu}$$

where $\mathbf{z} = (z_1, z_2)^{\mathrm{T}}$, $\boldsymbol{\nu} = (\xi_1, \xi_2)^{\mathrm{T}}$ and $\mathbf{H}$ is a 2×2 unit matrix.

Example 6.2. An equation of motion of a field controlled d.c. motor (Fig. 6.4).

Denote by Ω the armature velocity, by J the moment of inertia of the rotating parts, by M_r the damping torque, by $M = ki_a$ the developed torque, and by $C\Omega$ the moment of viscous friction. The

equation of motion for the motor shaft

$$J\dot{\Omega} = ki_a - C\Omega - M_r$$

The armature circuit is governed by

$$u = i_a R_a + L_a \frac{di_a}{dt} + k_e \Omega$$

or

$$L_a \frac{di_a}{dt} = -k_e \Omega - i_a R_a + u$$

Denoting $\dot{\theta} = \Omega$ reduces these equations to the form

$$J\ddot{\theta} = -C\dot{\theta} + ki_a - M_r$$
$$L_a \frac{di_a}{dt} = -k_e \dot{\theta} - R_a t_a + u$$

Dividing the first equation by J and the second by L_a and denoting

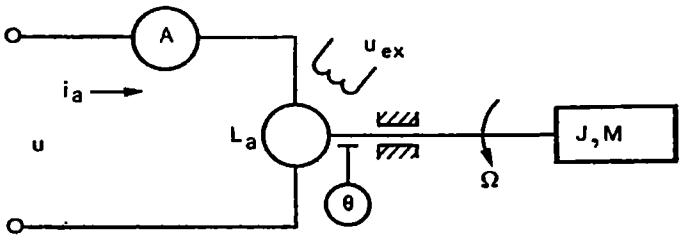


Fig. 6.4. Field controlled d.c. motor

$f_1 = -C/J$, $f_2 = k/J$, $\vartheta = -M_r/J$, $f_3 = -k_e/L_a$, $f_4 = -R_a/L_a$, and $a = 1/L_a$ gives

$$\ddot{\theta} = f_1\dot{\theta} + f_2 i_a + \vartheta$$
$$\frac{di_a}{dt} = f_3\dot{\theta} + f_4 i_a + au$$

Introducing the state variables $x_1 = \theta$, $x_2 = \dot{\theta} = \dot{x}_1$, $x_3 = i_a$, we obtain three first order differential equations

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = f_1 x_2 + f_2 x_3 + \vartheta$$
$$\dot{x}_3 = f_3 x_2 + f_4 x_3 + au$$

which may be written as one vector equation

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{G}u + \mathbf{\Gamma}\vartheta$$

where

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & f_1 & f_2 \\ 0 & f_3 & f_4 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

To monitor the operation of the motor assembly one may use an ampermeter A in the armature circuit, and a transducer of angle θ on the motor shaft. If we denote the indications of these devices by z_1 and z_2, and the measurement errors by ξ_1 and ξ_2, the measurement data may be written as

$$\begin{aligned} z_1 &= x_1 + \xi_1 \\ z_2 &= x_3 + \xi_2 \\ \text{or } \mathbf{z} &= \mathbf{H}\mathbf{x} + \boldsymbol{\nu} \end{aligned}$$

where

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\nu} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

6.3.5 Discrete Time Systems

When the system is controlled by a computer, the process control is effected in discrete equidistant time instants $t_n = nT_0$, where T_0 is the sampling period, assumed to be unity. Replacing in (6.33) the continuous $\mathbf{x}(t)$, $\mathbf{u}(t)$, $\vartheta(t)$, and $\dot{\mathbf{x}}(t) = d\mathbf{x}(t)/dt$ with the discrete $\mathbf{x}_n$, $\mathbf{u}_n$, ϑ_n, and $\Delta\mathbf{x}_n/T_0 = \Delta\mathbf{x}_n = \mathbf{x}_{n+1} - \mathbf{x}_n$, we write this differential equation as the difference equation

$$\mathbf{x}_{n+1} = \Phi[n]\,\mathbf{x}_n + \mathbf{G}[n]\,\mathbf{u}_n + \Gamma[n]\,\vartheta_n \tag{6.35}$$

where

$$\Phi[n] = 1 + \mathbf{F}[n] \tag{6.36}$$

The equation of measurements (6.34) now becomes

$$\mathbf{z}_n = \mathbf{H}\mathbf{x}_n + \boldsymbol{\nu}_n \tag{6.37}$$

6.4 Classification of Optimal Control Problems

6.4.1 One Step Decision Problems

One step decision problems are normally not concerned with the methods of implementation of a decision being made, i.e. they are aimed not at the magnitude and the character of the control action $\mathbf{u}$, but rather directly at the value of the state variable $\mathbf{x}$ that ensures the best path to the objective of control.

A one-step decision problem is assumed formulated if the given parameters include the space of states of nature Θ with the pro-

bability distribution density $\xi(\vartheta)$ for all ϑ in Θ, the solution space X, and a performance measure that in this case is usually called the *objective function*. Some authors prefer other terms for objective function, such as *payoff function* or *loss function* depending on applications.

The objective function, q, representing explicitly the objectives of control may be viewed as the output variable of the controlled system. By analogy with (6.26), a scalar objective function that depends on state of nature ϑ and state of the controlled system $\mathbf{x}$ may be represented as

$$q = q(\mathbf{x}, \vartheta) \tag{6.38}$$

Thus, a one-step decision problem is the triplet

$$G = (X, \Theta, q) \tag{6.39}$$

where q is a scalar function defined over the direct product of sets $X \times \Theta$.

Solution of this problem consists in finding $\mathbf{x}^*$ in X such that minimizes q, i.e. satisfies the condition

$$\mathbf{x}^* = \{\mathbf{x} \in X: q(\mathbf{x}, \vartheta) = \min\} \tag{6.40}$$

Note that if $q(\mathbf{x}, \vartheta)$ is a maximum at some $\mathbf{x}^*$, then $-q(\mathbf{x}, \vartheta)$ will be a minimum at the same $\mathbf{x}^*$. Therefore a minimization problem may readily be converted to the maximum type and vice versa.

A particular choice of a method for solving a one-step decision problem depends on the way in which the set of feasible solutions X has been specified, on the available data about the state of nature and on the type of the objective function q. A brief account of the basic solution techniques will be given below.

A problem is called *deterministic* if it involves no uncertainty with regard to the state of nature. In deterministic problems, the state space of nature Θ consists of only one element ϑ_0, the probability of which is unity. The appropriate objective function depends only upon the state of the controlled system

$$q = q(\mathbf{x}) = q(x^{(1)}, \ldots, x^{(n)}) \tag{6.41}$$

A one-step deterministic problem is said to be the *classical optimization problem* [39] if its constraints are of the type (6.24) and these constraints contain no inequalities, no non-negativity conditions or conditions of discreteness of the variables, $m < n$, and the functions $f_i(x^{(1)}, \ldots, x^{(n)})$ and $q(\mathbf{x})$ are continuous and at least twice differentiable. In this case the problem is formulated as follows.

Find $x^{(1)}, \ldots, x^{(n)}$ such that satisfy Eqs. (6.42) and minimize the function $q(x^{(1)}, \ldots, x^{(n)})$ subject to the constraints

$$f_i(x^{(1)}, \ldots, x^{(n)}) = 0, \quad i = 1, \ldots, m; \quad m < n \tag{6.42}$$

The most popular methods of one-step problem solving are collectively known as *mathematical programming*. These methods are efficient tools of evaluating the values of $x^{(1)}, \ldots, x^{(n)}$ minimizing the objective function q (x) subject to constraints of the type (6.24) in the form of equations and inequalities. The variables usually must be non-negative. It is worth noting that mathematical programming is an algorithmic rather than analytical form of problem solving, i.e. points out a computational procedure rather than provides a formula of the final result. Therefore, mathematical programming techniques become efficient predominantly where computers are available.

The simplest case of mathematical programming is *linear programming* (LP). In an LP problem the left-hand sides of the constraints (6.24) and the objective function (6.41) are both linear functions of $x^{(1)}, \ldots, x^{(n)}$. The problem is to find non-negative values of variables $x^{(1)}, \ldots, x^{(n)}$ such that minimize the objective function

$$q(x^{(1)}, \ldots, x^{(n)}) = \sum_j c_j x^{(j)} \tag{6.43}$$

ubject to the constraints

$$\sum_j a_{ij} x^{(j)} \leqslant b_i, \quad i = 1, \ldots, m. \tag{6.44}$$

Any other problem of mathematical programming is a problem of *nonlinear programming* (NLP). In nonlinear programming problems, either the objective function (6.41) or the left-hand sides of the constraints (6.24), or both are nonlinear functions of $x^{(1)}, \ldots, x^{(n)}$. The problem with the objective function and the constraints of the type (6.43) and (6.44) and the variables of the integer type also falls in the realm of nonlinear programming with the name of *integer programming* problem.

A one-step decision problem is said to be *stochastic* if the state space of nature Θ consists of more than one element, so that the known parameter is the probability distribution density over the space Θ rather than the real state of nature Θ.

Stochastic problems that require that the values of variables should be found such that satisfy the constraints (6.24) and minimize the objective function (6.38) are referred to as stochastic programming problems. In many cases, however, a somewhat different definition of the objective function can reduce the stochastic programming problem to the one of linear or nonlinear programming. Indeed, because the state of the world ϑ is a random variable with a probability distribution $\xi(\vartheta)$ over the space Θ, then the value of q (x, ϑ) at a given $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^{\mathrm{T}}$ will also be a random variable with the same probability distribution $\xi(\vartheta)$ over the space Θ. It would be

reasonable therefore to use the expectation of q (x, ϑ) as the objective function.

Hence, for random processes, according to (5.52) the objective function may be written as

$$q_1(\mathbf{x}) = \sum_{\vartheta \in \Theta} \xi(\vartheta)\, q(\mathbf{x}, \vartheta) \tag{6.45}$$

Because q_1 (x) is a deterministic function of x, the evaluation of $x^{(1)}, \ldots, x^{(n)}$ minimizing the objective function (6.45) subject to the constraints (6.24) may be effected on the lines of linear or nonlinear programming.

An important case of the one-step stochastic problem of decision making is the one of the quantities x_i, $i = 1, \ldots, n$, taking on only a finite set of values. The methods of solution for such problems fall into the realm of *statistical decision theory.*

Another class of problems covering an important field of applications consists of those where decision is made by several sides, say two, rather than one decision maker and the aims of these sides are in conflict. An example may be the hunting problem in which the distance between the hunted and the hunter depends on decisions and actions of both sides concerned. The hunter tends to cut this distance as much as possible, while the hunted would like to increase it. Solution techniques for such problems are discussed in game theory, where the persons responsible for decisions are referred to as *players.*

In a conflict situation, each player makes a decision independently of his opponent, therefore the mathematical description of such a situation treats the space of decisions as a direct product of two sets $X \times Y$, where $X = \{x_1, \ldots, x_n\}$ is the decision space of one player and $Y = \{y_1, \ldots, y_m\}$ is the decision space of the other player.

Elements of the decision space $X \times Y$ will be pairs (x, y), x in X and y in Y, i.e., will be governed by decisions taken by player I and player II from their own information sets. Assume for simplicity that there is no uncertainty in the state of nature, then the objective function

$$q = q(\mathbf{x}, \mathbf{y}) \tag{6.46}$$

depends only on elements of space $X \times Y$. The conflict of interests of the two players manifests itself in that the first player making his choice from the information set X tends to minimize the objective function by his option, while the second player choosing from Y tends to minimize it. The essence of this situation is that each player should make the best decision on his part keeping in mind that his opponent will also choose at his best.

6.4.2 Dynamic Optimization Problems

Among the many and diverse problems of cybernetics a considerable place is occupied by those in which the controlled system is continuously moving and changing under the influence of various external and internal factors. Control of such dynamic systems relates to the class of *dynamic control problems.*

The system is said to be *controllable* if among the factors that act on it there are such whose variation can be used to alter the system behavior. As will be recalled such targetted actions are called control actions, $\mathbf{u}(t)$.

The character of motion of the plant is defined by the system of differential equations (6.28) that may be conveniently represented in vector form (6.32). The control $\mathbf{u}(t)$ enters equations (6.32), therefore this equation defines not just a specific motion of the system but rather its technical possibilities that may be implemented by using one or another control from the space of feasible control options, U.

The analyst may assess how close a control brings the system to the desired objective by introducing, as before, an objective function of the type (6.26), which in this case is convenient to write as

$$q = q_\vartheta [\mathbf{x}(t), \mathbf{u}(t)] \tag{6.47}$$

By way of illustration, if $u(t)$ is a fuel flow rate, and $x(t)$ the plane velocity at this time instant, then from the standpoint of fuel consumption the performance measure for control at any time may be characterized by the quantity $q(t) = u(t)/x(t)$, i.e. by fuel consumed per unit length of flight, which, quite naturally, will depend on the state of nature ϑ, i.e. on the set of external factors governing the conditions of the flight.

The objective function (6.47) is seldomly used, as it is an instantaneous estimator of flight variables, whereas in the majority of practical situations it is required to track the system processes over the whole time of control from 0 to T.

In many applications the objective function can be selected such that the estimation of a process in the controlled system can be carried out by integrating the objective function over the period of control, i.e. the performance measure may be chosen as the functional

$$J(\mathbf{u}) = \int_0^T q_\vartheta [\mathbf{x}(t), \mathbf{u}(t)]\, dt \tag{6.48}$$

If the objective function has the meaning of losses, then equation (6.48) gives the total losses over the whole period of control.

Sometimes the objective of control is expressed as the desired behavior $\mathbf{z}(t)$ of the process under control. The appropriate objective

function then may be the square or absolute value of the deviation of the process $\mathbf{x}(t)$ from the desired behavior, viz.,

$$\begin{aligned} q &= [\mathbf{x}(t) - \mathbf{z}(t)]^2 \\ q &= |\mathbf{x}(t) - \mathbf{z}(t)| \end{aligned} \tag{6.49}$$

The respective performance index (6.48) will then be the square or absolute error.

In addition to constraints of the type (6.25) defining the space of feasible control actions, U, dynamic control problems are subject to integral constraints of the form

$$\int_0^T H_\vartheta [\mathbf{x}(t), \mathbf{u}(t)]\, dt \leqslant K = \text{const} \tag{6.50}$$

In some situations it is required to limit the variation range of some parameter $a(\mathbf{x}, \mathbf{u})$ in the process of control. Denote the upper bound of this range by a_0. If the integrand $H(\mathbf{x}, \mathbf{u})$, called the *penalty function*, is defined as

$$\begin{aligned} H(\mathbf{x}, \mathbf{u}) &= 0, && \text{for } a(\mathbf{x}, \mathbf{u}) \leqslant a_0 \\ &= a(\mathbf{x}, \mathbf{u}) - a_0]^2 && \text{for } a(\mathbf{x}, \mathbf{u}) > a_0 \end{aligned} \tag{6.51}$$

then the integral constraint (6.50) will express a requirement that a should not exceed a_0 except for a short time and then for a small amount only. This condition will be more severe the smaller the value of K in (6.50). At $K = 0$, this constraint does not allow any excess of a over a_0.

Constraints of the type (6.50) occur also where resources are limited, for example, limited amounts of energy or fuel where trajectories are concerned.

This preliminary discussion leads us to a definition of optimal control in dynamic systems. For the system described by the differential equation (6.32), an *optimal control*, $\mathbf{u}^*(t)$, selected from the space of feasible controls U, is such that minimizes the performance measure (6.48) subject to the given constraints (6.50) imposed on the available resources.

Like one-step control problems, dynamic problems may be *deterministic*, if the space of states of nature Θ consists of only one element $\vartheta = \vartheta_0$, and *stochastic*, if the state space Θ contains more than one element given by an *a priori* probability distribution density $\xi(\vartheta)$ on Θ.

An important class of stochastic problems are those of *adaptive control* that is employed when *a priori* data on the state of nature is insufficient for efficient management, or when a sufficiently accurate mathematical description of the controlled system is not available. Adaptive control aims at refining the knowledge on the state of

nature or the properties of the system in the process of control by way of trying various methods of control for the best one under the specific circumstances [34].

6.4.3 Terminal State Control

For some applications the behavior of the plant in the process of control is not of essential interest, and the thing that matters is the final state of the controlled system. Examples of such problems may be instrument landing of an aircraft that imposes strict requirements on the velocity and position at the instant it touches the ground, delivery of goods by certain terms to a given destination, or attainment to the end of a year a given labor productivity. These problems are referred to as ones of *terminal state control.*

Denote by $\mathbf{x}(T)$ the state of the plant at the final instant of time. The objective function is

$$q = q_0[\mathbf{x}(T)] \tag{6.52}$$

Because $\mathbf{x}(T)$ depends on the type of control used, the value of q will be also dependent on this control. Accordingly, the problem of choosing an optimal control may be formulated now as follows. For the plant described by the differential equation (6.32), from the space of feasible control actions U choose a control $\mathbf{u}^*(t)$ such that minimizes the objective function (6.52) subject to the constraints (6.50) imposed on the available resources.

6.5 Sequential Control

6.5.1 Behavior of the Dynamic System as a Function of Its Initial State

For many dynamic systems, evaluation of an optimal control becomes much easier if the process of control allows a natural or artificial partition into separate stages or steps. In order to consider this process in the general form we assume that the state of the system is described by the multidimensional variable $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^{\mathrm{T}}$.

If we assume that the process is not controlled and there is no uncertainty in the state of nature, the differential equation governing the system motion may be written by analogy with (6.32) as

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}), \quad \mathbf{x}(0) = \mathbf{c} \tag{6.53}$$

The solution of this equation is a function of time, $\mathbf{x} = \mathbf{x}(t)$. It is important to stress, however, that the solution of this equation depends on the initial state $\mathbf{c}$. We select a representation of this

solution such that would indicate its explicit dependence on both time and initial state, namely,

$$\mathbf{x} = \mathbf{x}(\mathbf{c}, t) = \mathbf{x}[\mathbf{x}(0), t] \tag{6.54}$$

This representation allows us to view the system state at any time instant t as a certain transformation of the initial state $\mathbf{x}(0) = \mathbf{c}$ on the interval t.

Consider the motion of a plant on the interval from 0 to t_2 partitioned by an interim point t_1 into two intervals t_1 and $\tau = t_2 - t_1$,

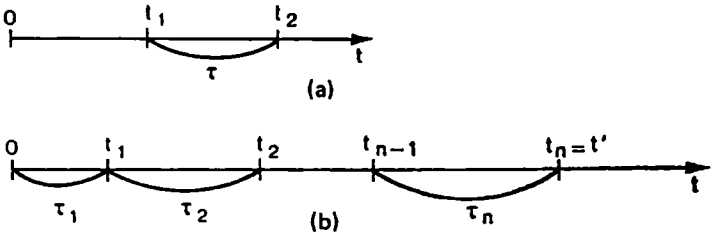


Fig. 6.5. Partition of an interval into (*a*) two and (*b*) n steps

as shown in Fig. 6.5*a*. This partition leaves us with three states, namely, the initial state $\mathbf{x}(0) = \mathbf{c}$, $\mathbf{x}(\mathbf{c}, t_1)$, and the terminal state $\mathbf{x}(\mathbf{c}, t_2)$. The last state may be described in two ways. It may be treated either as the transformation of the initial state $\mathbf{x}(0) = \mathbf{c}$ in the interval $t_2 = t_1 + \tau$, viz.,

$$\mathbf{x}(\mathbf{c}, t_2) = \mathbf{x}(\mathbf{c}, t_1 + \tau) \tag{6.55}$$

or as a transformation of state $\mathbf{x}(\mathbf{c}, t_1)$ in the interval τ

$$\mathbf{x}(\mathbf{c}, t_2) = \mathbf{x}[\mathbf{x}(\mathbf{c}, t_1), \tau] \tag{6.56}$$

Since both expressions describe one and the same state, then by equating them we get

$$\mathbf{x}(\mathbf{c}, t_1 + \tau) = \mathbf{x}[\mathbf{x}(\mathbf{c}, t_1), \tau] \tag{6.57}$$

6.5.2 Representing a Dynamic Process as a Sequence of Transformations

Assume that in the interval $[0, t']$ a dynamic process $\mathbf{x}(\mathbf{c}, t)$ can be represented, naturally or artificially, as a sequential, multiple-step control, and find an appropriate description for such a process. To arrive at a multi-step process we partition the interval $[0, t']$ into n steps of duration $\tau_1, \tau_2, \ldots, \tau_n$, as shown in Fig. 6.5*b*. Denote by t_k $(k = 0, \ldots, n)$ the end points of kth step, so that $t_{k+1} = t_k + \tau_{k+1}$, and by $\mathbf{x}_k$ the state of the system at instant t_k

$$\mathbf{x}_k = \mathbf{x}(\mathbf{c}, t_k) \tag{6.58}$$

Consider the state

$$\mathbf{x}_{k+1} = \mathbf{x}(\mathbf{c}, t_{k+1}) = \mathbf{x}(\mathbf{c}, t_k + \tau_{k+1}) \tag{6.59}$$

In agreement with (6.57) and (6.58) this expression may be written as

$$\mathbf{x}_{k+1} = \mathbf{x}[\mathbf{x}(\mathbf{c}, t_k), \tau_{k+1}] = \mathbf{x}(\mathbf{x}_k, \tau_{k+1}) \tag{6.60}$$

This expression gives the state $\mathbf{x}_{k+1}$ of the system as the result of transformation of state $\mathbf{x}_k$ at step $k+1$.

Introduce an operator T that represents the transformation of system state in one step, namely,

$$\mathrm{T}(\mathbf{x}_k) = \mathbf{x}(\mathbf{x}_k, \tau_{k+1}), \quad k = 0, \ldots, n-1 \tag{6.61}$$

Now the relationship (6.60) rewrites as

$$\mathbf{x}_{k+1} = \mathrm{T}(\mathbf{x}_k) \tag{6.62}$$

Letting $k = 0, 1, \ldots, n-1$, we may describe the whole of the dynamic process as a sequence of transformations

$$\mathbf{x}_0 = \mathbf{c}, \ \mathbf{x}_1 = \mathrm{T}(\mathbf{x}_0), \ \ldots, \ \mathbf{x}_n = \mathrm{T}(\mathbf{x}_{n-1}) \tag{6.63}$$

6.5.3 Sequential Control Process

A dynamic process described by the transformation (6.62) is uncontrollable. In order to make it controllable one needs to carry out at each step one of several transformations $\mathrm{T}_1(\mathbf{x}_k), \ldots, \mathrm{T}_r(\mathbf{x}_k)$ rather than one transformation.

It is convenient to assume that the particular form of transformation will depend on the parameter $\mathbf{u}_k$ which may take on one value of the set U_k at the kth step. We shall call this parameter a control, and the set U_k the space of feasible controls at the kth step. Now, the transformation performed at step k may be written as

$$\mathbf{x}_{k+1} = \mathrm{T}(\mathbf{x}_k, \mathbf{u}_k), \ \mathbf{u}_k \in U_k \tag{6.64}$$

If we let in succession in (6.64) $k = 0, 1, \ldots, n-1$, and observe the initial state $\mathbf{x}_0$, then we obtain a description of the entire controlled multi-step process

$$\mathbf{x}_{k+1} = \mathrm{T}(\mathbf{x}_k, \mathbf{u}_k), \quad \mathbf{u}_k \in U_k, \quad \mathbf{x}_0 = \mathbf{x}(0) = \mathbf{c}, \quad k = 0, 1, \ldots, n-1 \tag{6.65}$$

This expression is called the difference equation of the controlled plant. It is similar to the differential equation (6.32) describing continuous dynamic processes.

Example 6.3. Assume that the state variable is a two-dimensional variate $\mathbf{x} = (x^{(1)}, x^{(2)})$ and that it may take on values defined by the nodes of the mesh shown in Fig. 6.6*a*. Motion from one node to

the next is effected by one of two controls: $u_k = 0$ allowing horizontal motion, and $u_k = 1$, vertical motion. Consequently, the space of feasible controls is identical for any step

$$U_k = \{0, 1\}, \quad k = 0, 1, \ldots, n-1$$

Consider one of the cells of the mesh displayed in Fig. 6.6*b*. Assume that the system finds itself in the left bottom corner of the cell after

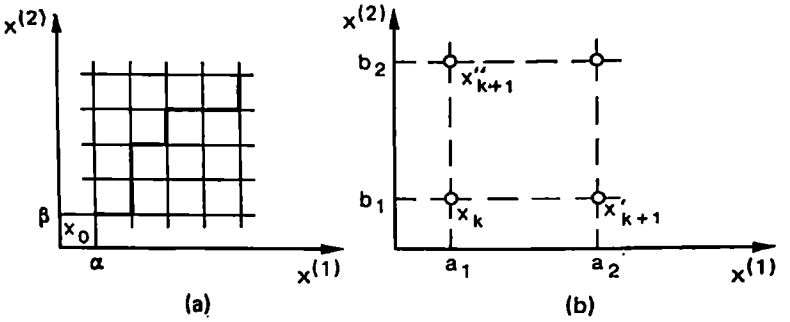


Fig. 6.6. Multi-step process with two-dimensional variable

the kth step, so that $\mathbf{x}_k = (a_1, b_1)$. The value $\mathbf{x}_{k+1} = \mathrm{T}(\mathbf{x}_k, u_k)$ depends on the control used. The following relations may be written

$$\mathrm{T}[(a_1, b_1), 0] = (a_2, b_1) = x'_{k+1}$$
$$\mathrm{T}[(a_1, b_1), 1] = (a_1, b_2) = \mathbf{x}''_{k+1}$$

A particular trajectory of system motion may be described by indicating the initial state $\mathbf{x}_0$ and the sequence of controls used. The trajectory indicated by a heavy line in Fig. 6.6*a* is the result of the control sequence $\mathbf{u} = (01101001)$ applied at the initial conditions $\mathbf{x}_0 = (\alpha, \beta)$.

6.5.4 Performance Measure

We assume that the effectiveness of control is defined by the value of the objective function q. Its numerical value may be viewed as losses inflicted by a particular control. The loss per step will depend on the state of the process at the outset of the particular step and on the control used at this step, i.e.

$$q_k = q(\mathbf{x}_k, \mathbf{u}_k), \quad \mathbf{u}_k \in U_k \tag{6.66}$$

As a performance criterion we may take the losses resulted after all n steps of the process and represent this performance index in the form

$$J_n(\mathbf{u}) = \sum_{k=0}^{n-1} q(\mathbf{x}_k, \mathbf{u}_k) \tag{6.67}$$

Here, $\mathbf{u}$ denotes the sequence of control actions, i.e. an ordered set of the form

$$\mathbf{u} = (\mathbf{u}_0, \ldots, \mathbf{u}_{n-1})^{\mathrm{T}}, \quad \mathbf{u}_0 \in U_0, \ldots, \mathbf{u}_{n-1} \in U_{n-1} \tag{6.68}$$

If we consider a multistep process of terminal-value control, then the losses will depend on only the system state at the end of the process, which in turn depends on the initial state and the controls used at each step

$$q = q(\mathbf{x}_n) = q(\mathbf{x}_0, \mathbf{u}) \tag{6.69}$$

where $\mathbf{u}$ is given by (6.68). The scalar function (6.69) is the objective function of the sequential control process.

The problem of evaluation of optimal control under a sequential process may be formulated as follows. Given a dynamic system whose process is described by the difference equation (6.65), find a sequence of control actions $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$ minimizing the performance measure (6.67) or the objective function (6.69) subject to the constraints $\mathbf{u}_k \in U_k$, $k = 0, 1, \ldots, n-1$.

Because sequential multi-step control processes are a particular case of dynamic control processes, they embrace all those types of problem that were considered under dynamic control problems. For example, if the space of states of nature Θ consists of one element ϑ_0, then the multi-step control problem is called *deterministic*, otherwise it is referred to stochastic problems.

It should be noted that a problem may not always be uniquely related to sequential or one-step problems. For example, if the sequence of values of a variable in a sequential problem is viewed as a single multivariate, then the sequential problem becomes a one-step problem.

6.6 Basic Concepts of Optimization

6.6.1 General Statement of Optimization Problem

The general optimization problem seeks for a vector $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^{\mathrm{T}}$ from the feasible domain X such that minimizes the objective function $q(\mathbf{x})$, i.e. a vector $\mathbf{x}^*$ in X for which

$$q(\mathbf{x}^*) \leqslant q(\mathbf{x}), \qquad \mathbf{x} \in X \tag{6.70}$$

If such a vector exists, then it defines a *weak global (overall) minimum* of $q^*(\mathbf{x})$ in the domain X. This minimum is called *weak* because it satisfies the "weak" inequality, and *global* or *overall* because the inequality holds for any $\mathbf{x}$ in X. A minimum at $\mathbf{x} = \mathbf{x}^*$ is called *strong* if $q(\mathbf{x}^*) < q(\mathbf{x})$ for $\mathbf{x} \neq \mathbf{x}^*$. Changing the signs in these inequalities, we obtain, respectively a *weak* and a *strong* maximum. However, a maximum of $q(\mathbf{x})$ gives a minimum of

— q (x), therefore in what follows we shall confine ourselves to the minimization problem.

A weak global minimum implies that the optimal point may not be unique as any x satisfying the equality q (x) = q (x*) is also an optimal point. A strong global minimum is always unique.

Although the aim of any optimization procedure is attainment of a global minimum of the objective function, the concept of local or relative minimum is of significance for this problem.

A minimum at x = x* is called a *local (relative) minimum* if there exists a neighborhood Q_ε (x*) of point x* such that for all x in Q_ε (x*) q (x*) $\leqslant q$ (x).

If q (x) is differentiable then the search for local minima reduces to the evaluation of stationary points at which the partial derivatives of q (x) vanish,

$$\frac{\partial q(\mathbf{x})}{\partial x^{(i)}} = 0, \quad i = 1, \ldots, n \tag{6.71}$$

It is worth noting that (6.71) is a necessary but not sufficient condition of existence of a relative minimum. A stationary point may be a point of relative maximum and a saddle point at which neither relative minimum nor relative maximum is achieved. As has been indicated in Sect. 4.7 a sufficient condition of a relative minimum at a stationary point is that the quadratic form (4.111) be positive definite.

Although an overall minimum is at the same time a relative minimum, the converse is not necessarily true. Moreover, on a feasible domain X the function q (x) may have not one but a few relative minima. At the same time, most of the known optimization techniques reveal only the points of relative minimum. Therefore, there remains a very complicated problem of evaluating such an absolute minimum that is global.

6.6.2 Constraints

A typical method of definition of a feasible set X is to impose constraints on the variables as in (6.24) and require also that all or some of the variables be non-negative. If the constraints are more than one, then the feasible set of solutions of the optimization problem is the intersection of all sets defined by these constraints.

For a two-dimensional variable x = (u, v), two constraints

$$u^2 + v^2 \leqslant r^2 \text{ and } uv \geqslant c$$

are shown in Fig. 6.7*a*. The first of them defines a circle of radius r centered on the origin, the second is a region separated from the origin by two branches of a hyperbola. In this case the feasible region is defined by unhatched areas in the first and third quadrants.

If in addition it is required that the variables be non-negative, then there remains only the area in the first quadrant.

Figure 6.7*b* shows the feasible region defined by the constraints

$$u^2 + v^2 \leqslant r^2,\ u \geqslant 0,\ v \geqslant 0$$

With some types of constraints the feasible set may be empty. For example, if we reduce the radius of the circle in Fig. 6.7*a* then there

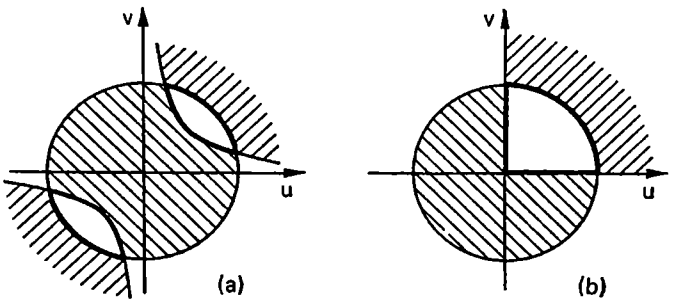


Fig. 6.7. Admissible sets

will be no points in common inside the circle and in the region defined by the branches of the hyperbola. Such constraints are referred to as *inconsistent*.

Many optimization techniques considered below are applicable only if the feasible set is *convex*. Assuming the nonnegativity of the variables, the feasible sets shown in Fig. 6.7 are convex. However, if we change the inequality sign for the circle in Fig. 6.7*b*, then the feasible set becomes the first quadrant less the unhatched sector and will not be convex any longer.

Generally speaking, the optimization problem does not always have a solution. For instance, the problem of max $(u + v)$ subject to $u \leqslant v$ has no solution because for any $u + v$ other u and v can be found that make a greater sum. Nevertheless a wide class of problems can be identified for which the existence of an optimum can be guaranteed by the following theorem.

Theorem of Weierstrass. A continuous function defined on a nonempty, bounded, closed set has a minimum (maximum) in at least one point of this set.

It is important therefore for most problems of optimization that the feasible set should be closed. To ensure this property all the constraints should have $\leqslant$ or $\geqslant$ signs, where the equality corresponds to the boundary points of the feasible set. The Theorem of Weierstrass gives sufficient conditions of optimality, however, problems where its conditions are not satisfied may also have solutions.

6.6.3 Convex and Concave Functions

Most of available solution techniques for optimization problems reduce to the investigation of the behavior of q (x) in the neighborhood of x, i.e. to the investigation whether or not x is a relative solution (minimum or maximum). The problem becomes much more complicated if the objective function q (x) has several maxima or minima in the feasible region X. We shall be interested therefore in problems where the objective function has only one maximum or minimum and where, consequently, the relative solution is global, too. The concepts of concave and convex function are of significance for such problem formulation.

Let f (x) be a function given on a convex set X, $\mathbf{x}_1$ and $\mathbf{x}_2$ be two arbitrary points in X, and $\hat{\mathbf{x}} = \lambda \mathbf{x}_1 + (1 - \lambda)\, \mathbf{x}_2$ for $0 \leqslant \lambda \leqslant 1$ be an

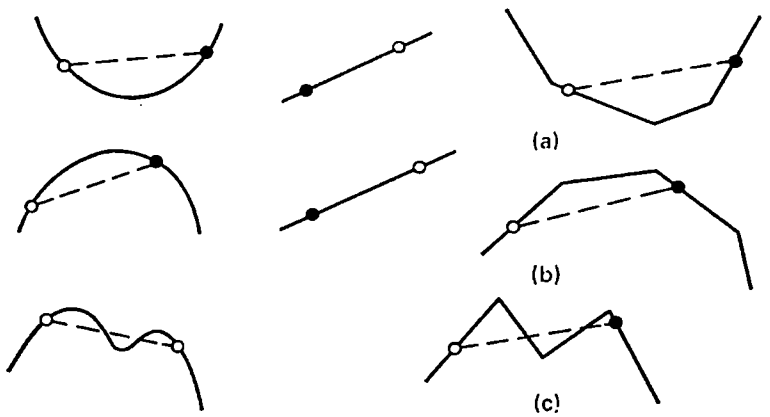


Fig. 6.8. Types of functions

arbitrary point in the line segment connecting $\mathbf{x}_1$ and $\mathbf{x}_2$. Consider also the line segment $z = \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2)$ connecting the values $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$.

The function f (x) is called *convex* if its values between any two points lie below (not above) the line segment connecting these two points, i.e. if for any $\mathbf{x}_1$ and $\mathbf{x}_2$ and any $0 \leqslant \lambda \leqslant 1$ f (x) will not exceed values z of the line segment connecting $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$, namely,

$$f[\lambda \mathbf{x}_1 + (1 - \lambda)\, \mathbf{x}_2] \leqslant \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2) \qquad (6.72)$$

A function is called *concave* if its values between any two points lie entirely above (not below) the line segment connecting these two points, i.e. if

$$f[\lambda \mathbf{x}_1 + (1 - \lambda)\, \mathbf{x}_2] \geqslant \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2) \qquad (6.73)$$

Geometrical interpretations of (*a*) convex, (*b*) concave, and (*c*) neither concave nor convex functions are shown in Fig. 6.8.

If we change the signs $\leqslant$ and $\geqslant$ in (6.72) and (6.73) for the signs $<$ and $>$ of strict inequalities, then we arrive at a definition of a strictly convex and a strictly concave function.

Theorem 6.1. The sum of convex (concave) functions is also a convex (concave) function.

Proof. Consider the function $f(\mathbf{x}) = \Sigma f_i(\mathbf{x})$, where $f_i(\mathbf{x})$ are convex functions on a convex set X. For arbitrary $\mathbf{x}_1$ and $\mathbf{x}_2$ in X and any $0 \leqslant \lambda \leqslant 1$

$$\begin{aligned} f[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] &= \sum_i f_i[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \\ &\leqslant \sum_i [\lambda f_i(\mathbf{x}_1) + (1-\lambda) f_i(\mathbf{x}_2)] \\ &= \lambda \sum_i f_i(\mathbf{x}_1) + (1-\lambda) \sum_i f_i(\mathbf{x}_2) \\ &= \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \end{aligned}$$

i.e. the condition for $f(\mathbf{x})$ to be convex is satisfied. A similar argument may be used to prove the theorem for concave functions.

Consider some important types of convex and concave functions. The linear function $\mathbf{z} = \mathbf{ax}$ is convex (and concave) in $R^{(n)}$ since for any $\mathbf{x}_1$ and $\mathbf{x}_2$ and any λ

$$\mathbf{a}[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] = \lambda \mathbf{a}\mathbf{x}_1 + (1-\lambda)\mathbf{a}\mathbf{x}_2$$

However, the linear function is neither strictly convex nor strictly concave.

The positive (nonnegative) definite quadratic form $\mathbf{z} = \mathbf{x}^T\mathbf{A}\mathbf{x}$ is a convex function in the entire space $R^{(n)}$. Indeed, for any $\mathbf{x}_1$ and $\mathbf{x}_2$ in $R^{(n)}$ and $\hat{\mathbf{x}} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$,

$$\begin{aligned} \hat{\mathbf{x}}^T\mathbf{A}\hat{\mathbf{x}} &= [\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2]^T\mathbf{A}[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \\ &= [x_2 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)]^T\mathbf{A}[\mathbf{x}_2 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)] \\ &= \mathbf{x}_2^T\mathbf{A}\mathbf{x}_2 + 2\lambda(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{A}\mathbf{x}_2 + \lambda^2(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2) \end{aligned}$$

Since by assumption $\mathbf{x}^T\mathbf{A}\mathbf{x} \geqslant 0$ and $0 \leqslant \lambda \leqslant 1$, then $\lambda \mathbf{x}^T\mathbf{A}\mathbf{x} \geqslant \lambda^2 \mathbf{x}^T\mathbf{A}\mathbf{x}$. Hence,

$$\begin{aligned} \hat{\mathbf{x}}^T\mathbf{A}\hat{\mathbf{x}} &\leqslant \mathbf{x}_2^T\mathbf{A}\mathbf{x}_2 + 2\lambda(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{A}\mathbf{x}_2 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_1 - \mathbf{x}_2) \\ &= \mathbf{x}_2^T\mathbf{A}\mathbf{x}_2 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{A}\mathbf{x}_2 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{A}\mathbf{x}_1 \end{aligned}$$

If we take into account that for a real-valued symmetric matrix $\mathbf{A}$

$$\mathbf{x}_1^T\mathbf{A}\mathbf{x}_2 = (\mathbf{x}_1^T\mathbf{A}\mathbf{x}_2)^T = \mathbf{x}_2^T\mathbf{A}^T\mathbf{x}_1 = \mathbf{x}_2^T\mathbf{A}\mathbf{x}_1$$

we obtain

$$\hat{\mathbf{x}}^{\mathrm{T}}\mathbf{A}\hat{\mathbf{x}} \leqslant \lambda \mathbf{x}_1^{\mathrm{T}}\mathbf{A}\mathbf{x}V_1 + (1 - \lambda)\,\mathbf{x}_2^{\mathrm{T}}\mathbf{A}\mathbf{x}_2$$

This implies that the function $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is convex.

A similar argument can be used to prove that a *negative (nonpositive) definite guadratic form is a concave function in all* R^n.

The sum of a linear function and the quadratic form $\mathbf{a}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is convex, if the quadratic form is positive definite, and it is concave if the quadratic form is negative definite. This conclusion is a corollary of Theorem 6.1.

6.6.4 Properties of Convex Functions

Here we give without proof some properties of convex functions that are important for NLP problems. Geometrical interpretations of these properties are shown in Fig. 6.9.

Let $f(\mathbf{x})$ be a convex function given on a closed convex set X. Then (1) any local minimum of $f(\mathbf{x})$ in X is global, (2) two and more

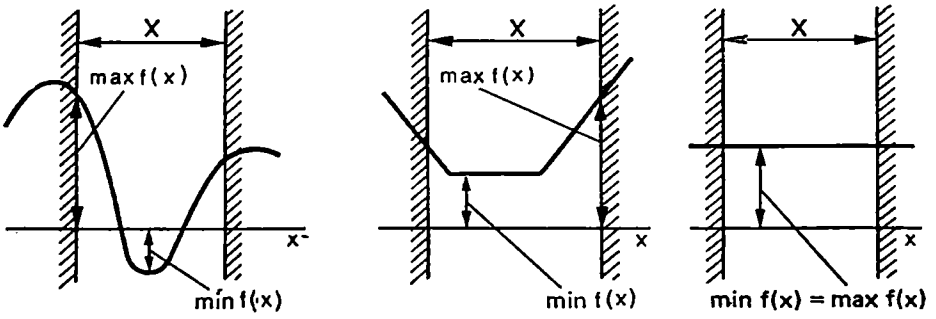


Fig. 6.9. Minima and maxima of convex functions

different local minima cannot exist, (3) if a global minimum is achieved at two distinct points, then it is achievable at infinite number of points, and (4) a global maximum of $f(\mathbf{x})$ is attainable at one or more extreme points of X.

Concave functions possess similar properties with respect to the maxima of $f(\mathbf{x})$.

6.6.5 Classical Optimization Problem

As has been pointed out the classical optimization problem is to find a minimum of the objective function $q(\mathbf{x})$, where $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^{\mathrm{T}}$ is a point in $R^{(n)}$, subject to the equality constraints

$$f_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m, \quad m < n \tag{6.74}$$

If the constraint qualification (6.74) holds, then the minimum of q (x) is said to be a *conditional minimum*. When these constraints are absent, we speak of an *unconditional minimum* whose evaluation reduces to the evaluation and investigation of stationary points of q (x).

The classical solution technique for this problem consists in using equations (6.74) for the elimination of m variables. The objective function is then reduced to the form

$$q\,(x^{(1)}, \ldots, x^{(n)}) = q_1\,(y^{(1)}, \ldots, y^{(n-m)}) \qquad (6.75)$$

where $y^{(1)}, \ldots, y^{(n-m)}$ are the remaining, renumbered variables. Now, the problem is to find the values of $y^{(1)}, \ldots, y^{(n-m)}$ that minimize q_1 under no constraints, i.e. the problem reduces to the unconstrained search for an unconditional extremum.

Example 6.4. Figure 6.10 shows a geometric interpretation of this problem formulated for the two-dimensional variable $\mathbf{x} = (u, v)^T$ subject to one constraint f (x) $= u^2 - v + 1 = 0$ and with the objective function q (x) $= (u - 1)^2 + v^2$. The graph shows the constraint qualification f (x) $= 0$ and lines q (x) $=$ const. The optimal solution x* is obtained at the point of tangency of two curves f (x) $= 0$ and q (x) $=$ const.

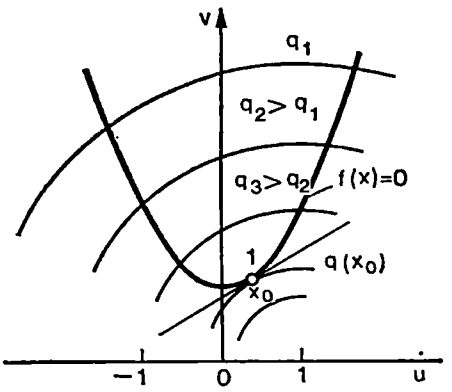


Fig. 6.10. Geometric interpretation of a conditional extremum problem

To obtain this solution analytically, we solve f (x) $= 0$ to get $v = u^2 + 1$. Substituting this expression into q (x), we get q (x) $= (u - 1)^2 + (u^2 + 1)^2$. Differentiating this expression with respect to u, we obtain the equation $2u^3 + 3u - 1 = 0$ that has only one real-valued root $u^* = 0.313$, which leads to $v^* = 1.098$ and q (x*) $= 1.678$.

It is instructive to note that lifting the constraint f (x) $= 0$, we obtain the minimum of q (x) at x $= (1, 0)$ and this minimum equals zero. Thus, the constraint results in a larger value of the objective function, i.e. impairs the quality of the optimum.

If equations (6.74) are complex, then the analyst is faced with considerable difficulties in eliminating m variables from q (x). A popular method to reduce the problem of conditional extremum to the problem of unconditional extremum is based on Lagrange multipliers.

6.6.6 Lagrange Multipliers Method

Consider a vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ and investigate the properties of the *Lagrangian function* defined as follows:

$$L(\mathbf{x}, \lambda) = q(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) \tag{6.76}$$

where the parameters λ_i are called the *Lagrange multipliers*. As can be seen, $L(\mathbf{x}, \lambda)$ is a function of $n + m$ variables $x^{(1)}, \ldots, x^{(n)}$, $\lambda_1, \ldots, \lambda_m$.

Let us investigate the stationary points of $L(\mathbf{x}, \lambda)$ which can be identified by applying the conditions

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda_j} = f_j(\mathbf{x}) = 0, \quad j = 1, \ldots, m \tag{6.77}$$

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial x^{(i)}} = 0, \quad i = 1, \ldots, n \tag{6.78}$$

It is easy to note that equations (6.77) coincide with the constraints (6.74) which implies that optimization of the function $q(\mathbf{x})$ subject to $f(\mathbf{x}) = 0$ is equivalent to the optimization of the Lagrangian function $L(\mathbf{x}, \lambda)$. Another way of saying it is that the conditional minimization problem of the objective function $q(\mathbf{x})$ subject to equality constraints boils down to the identification of stationary points of the Lagrangian function $L(\mathbf{x}, \lambda)$.

For the problem of Example 6.4, $L(\mathbf{x}, \lambda) = (u - 1)^2 + v^2 + \lambda(u^2 - v + 1)$. Differentiating this function with respect to u, v, and λ, we get three equations

$$\begin{aligned} u(1 + \lambda) - 1 &= 0 \\ 2v - \lambda &= 0 \\ u^2 - v + 1 &= 0 \end{aligned}$$

that can be solved for u^*, v^*, and λ.

7

LINEAR AND NONLINEAR PROGRAMMING

7.1 LP Problem Formulation

7.1.1 Basic Concepts and Definitions

In Chapter 6 we have learned that the term linear programming is associated with the analysis and solution of the following problem [43-45].

Given the system of m linearly independent equations in decision variables $x_1, \ldots, x_n$. As the name of this technique suggests these relations must all be linear,

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= b_1 \\ a_{m1}x_1 + \ldots + a_{mn}x_n &= b_m \end{aligned} \tag{7}$$

Without loss of generality we may assume also that $b_i \geqslant 0$, $i = 1, \ldots, m$.

A characteristic feature of this problem is that the number of these constraint equations is less than the number of unknowns, i.e. $m < n$. The problem is to find nonnegative values of x_i, $i = 1, \ldots, n$, that minimize the objective function,

$$q = c_1x_1 + \ldots + c_nx_n \tag{7.2}$$

subject to the constraint equations (7.1).

Typically the same linear programming problem can be expressed in a more condensed form using matrix notation. Minimize

$$q\,(\mathbf{x}) = \mathbf{cx} \tag{7.3}$$

subject to

$$\mathbf{Ax} = \mathbf{b}\ (\mathbf{b} \geqslant 0),\ \mathbf{x} \geqslant 0$$

where $\mathbf{A}$ represents an $m \times n$ matrix, $\mathbf{c}$ is a row n-vector, and $\mathbf{x}$ and $\mathbf{b}$ are column vectors.

Actually the system (7.1) is solved by matrix algebra when $m = n$. If the determinant of the system is nonzero, the system has a unique solution and can be solved by well developed techniques discussed in Sect. 4.6. When $m < n$, the system (7.1) has infinitely many solutions, i.e. there exists infinite number of sets of decision variables x_i that satisfy Eqs. (7.1). Each set of decision variables satisfying Eqs. (7.1) is thus a *solution* of the system.

In general there are infinitely many solutions that satisfy the additional restriction $x_i \geqslant 0$. These solutions will be called *feasible* or *admissible solutions*. The essence of the linear programming problem is to choose from the set of feasible solutions the one that minimizes the objective function (7.2).

Some LP problems (LPs) have all or some of relationships (7.1) in the form of inequalities. For example, in place of $a_{j1}x_1 + \ldots + a_{jn}x_n = b_j$ the system (7.1) may have $a_{j1}x_1 + \ldots + a_{jn}x_n \leqslant b_j$ or $a_{j1}x_1 + \ldots + a_{jn}x_n \geqslant b_j$. Then such inequalities can readily be made equations by introducing an additional, *slack variable* $x_{n+j} \geqslant 0$. This addition increases the number of variables without essentially altering the problem.

When the problem is to maximize, not minimize, the objective function (7.2), it may be reduced to the above by changing the sign of q. In what follows we shall encounter both types of problem and we denote by q the objective function to be minimized and by q' the one to be maximized.

Before we start to seek the solution of Eqs. (7.1) minimizing q subject to the restrictions $x_i \geqslant 0$, $i = 1, \ldots, n$, we try to find a solution of (7.1). Because the number of variables, n, in this system exceeds the number of equations, m, one of admissible solutions can be found by putting some $n - m$ of the variables to zero. The resulting system of m equations in m variables can be solved by common algebraic methods. It is important, of course, that the determinant of the coefficients would be nonzero. If this condition is not met, then we may equate to zero other $n - m$ variables. A solution obtained in this way is called a *basic solution*. The pertinent terminology of LP problems is given below.

Any set of m variables such that the determinant of the coefficients of these variables is nonzero is referred to as a *basis*. The m variables are said to be *in solution* while all others are zero. The variables in solution are also often called *basic variables* while those not in solution, i.e. equal to zero, are called *nonbasic variables*. Each system of equations (7.1) may have several different sets of basic variables.

If we put all nonbasic variables equal to zero and solve the resultant system of m equations in m unknowns, we obtain a *basic solution*. However, among the various basic solutions some will have negative values of some variables. Those solutions contradict the conditions and are therefore nonfeasible solutions.

A basic solution is feasible if it gives nonnegative values of the variables in solution. Feasible basic solutions are the simplest of the feasible solutions to the system (7.1). One additional condition is imposed on a problem solution, namely, the objective function (7.2) should be a minimum with this solution. This condition complicates the solution of the problem, however, the concept of a feasible basic solution is of significance in the evaluation of an optimal solution.

7.1.2 Working Simple LP Problems

Linear programming is a mathematical means that has been developed in assigning a fixed amount of resources to satisfy a number of demands in such a way that some objective is optimized and other defined conditions are still satisfied. In brief, it is required to find values of parameters such that would be best from some standpoint. These problems include optimization of the limited raw material supplies, making best use of the limited production facilities, efficient transportation schemes, and the like. We illustrate the essence of linear programming problems, or linear programs (LPs), by working specific examples.

Example 7.1. *Resource allocation problem.* To produce l different types of product $P_1, \ldots, P_l$, the manager has m types of resources $R_1, \ldots, R_m$ (raw materials, fuel, tool, and such) which are available in limited amount or number $b_1, \ldots, b_m$, respectively.

Let a_{ij} be the type R_i resource requirement per piece of type P_j product, and c_j be the profit per piece of type P_j product. For $l = 3$ and $m = 4$ these data are given in Table 7.1.

Table 7.1

Resource	Resource requirements			Stock
	P_1	P_2	P_3	
R_1	a_{11}	a_{12}	a_{13}	b_1
R_2	a_{21}	a_{22}	a_{23}	b_2
R_3	a_{31}	a_{32}	a_{33}	b_3
R_4	a_{41}	a_{42}	a_{43}	b_4
Profit per piece	c_1	c_2	c_3	—

The problem is to determine how many units of each product should be produced to maximize total profit.

Denote by x_j the number of pieces of type P_j. The restrictions in this problem are that the demand for resources of all items should not exceed the available stock

$$\sum_{j=1}^{l} a_{ij}x_j \leqslant b_i, \quad i = 1, \ldots, m \tag{7.4}$$

These constraints can readily be converted into equations by introducing slack variables $x_{l+i} \geqslant 0$ implying unused resources R_i. Then

$$\sum_{j=1}^{l} a_{ij}x_j + x_{l+i} = b_i, \quad i = 1, \ldots, m \tag{7.5}$$

Total profit is

$$q' = \sum_{j=1}^{l} c_j x_j \tag{7.6}$$

The optimal allocation will be nonnegative solution of the system (7.5) which maximizes the objective function (7.6).

Example 7.2. *Production scheduling problem.* According to a contract a factory is to produce in a time T l types of similar items, N_i pieces of each product. These items can be manufactured on r similar machine tools at a production rate a_{ij} items i on machine j, at a cost of b_{ij} monetary units per unit of product i.

If x_{ij} is the production item of machine j in manufacturing item i, then

$$\sum_{i=1}^{l} x_{ij} \leqslant T, \quad j = 1, \ldots, r \tag{7.7}$$

and

$$\sum_{j=1}^{r} a_{ij}x_{ij} = N_i, \quad i = 1, \ldots, l \tag{7.8}$$

The objective function will be the total cost of the whole production. Observing that product i manufactured on machine j costs $a_{ij}b_{ij}x_{ij}$, we have for the total costs of all manufactured items

$$q = \sum_{j=i=1}^{r} \sum^{l} a_{ij}b_{ij}x_{ij} \tag{7.9}$$

It is required to find x_{ij} such that minimize this quantity subject to the constraints (7.7) and (7.8).

Example 7.3. *Transportation problem.* The transportation problem requires the allocation of units or goods, available in amounts a_i, at $i = 1, \ldots, l$ origins, to a number of destinations, where the amount required at each destination is b_j, $j = 1, \ldots, r$, in such a way that the allocation is an optimum (least cost or maximum profit). It is assumed that the total demand equals the available stock, namely,

$$\sum_{i=1}^{l} a_i = \sum_{j=1}^{r} b_j \tag{7.10}$$

If we denote by x_{ij} the amount allocated from origin i to destination j, and by c_{ij} the cost or profit of allocating one unit from origin i

to destination j, then

$$\sum_{j=1}^{r} x_{ij} = a_i, \quad i = 1, \ldots, l \tag{7.11}$$

$$\sum_{i=1}^{l} x_{ij} = b_j, \quad j = 1, \ldots, r \tag{7.12}$$

The amounts available at each origin and the amounts required at each destination are often referred to as *rim requirements.*

The objective function defines the total cost (profit) of all allocations,

$$q = \sum_{j=1}^{r} \sum_{i=1}^{l} c_{ij} x_{ij} \tag{7.13}$$

It is instructive to note that the problem formulation remains the same where the amount of available goods exceeds the total demand. This situation is handled by introducing a fictitious destination $r + 1$ to which the surplus is allocated at a cost of $c_{i,\, r+1} = 0$, $i = 1, \ldots, l$.

Example 7.4. *Choice of an optimal system version.* It is required to select an optimum version of available versions of a computing system designed to perform in tandem r operations in r consequtive modules. There are l options for each module, say built around electronic tubes, semiconductor elements, ferrite-transistor elements, microminiature modules, and the like. The constraints are imposed on the system cost, X, size, Y, and speed, Z.

Denote by x_i, y_i, and z_i the cost, size and time required per one operation for module i, then the constraint may be written as

$$\sum_{i=1}^{r} x_i \leqslant X, \quad \sum_{i=1}^{r} y_i \leqslant Y, \quad \sum_{i=1}^{r} z_i \leqslant Z \tag{7.14}$$

The quantities x_i, y_i, and z_i depend on what version will be selected to make a particular module, hence, are the elements of the following sets:

$$\begin{aligned} x_i &\in \{x_{i1}, \ldots, x_{il}\} \\ y_i &\in \{y_{i1}, \ldots, y_{il}\} \\ z_i &\in \{z_{i1}, \ldots, z_{il}\} \end{aligned} \tag{7.15}$$

where x_{ij}, y_{ij}, and z_{ij} are the cost, size, and operation time of module i made in version j.

If c_1, c_2, and c_3 are the coefficients representing relative losses involved in reducing costs, size, and operating time, then the con-

dition for an optmum version may be written as

$$q = c_1 \sum_{i=1}^{r} x_i + c_2 \sum_{i=1}^{r} y_i + c_3 \sum_{i=1}^{r} z_i = \min \tag{7.16}$$

This problem differs from the LP problem formulated earlier in that its variables x_i, y_i, and z_i take on only the values of the finite sets (7.15) rather than any nonnegative values as before. Therefore, it should be handled by the methods of *integer programming*.

7.1.3 Geometrical Interpretation

Consider an optimization problem given by the system equations

$$\begin{aligned} -2x_1 + x_2 + x_3 &= 2 \\ x_1 - 2x_2 + x_4 &= 2 \\ x_1 + x_2 + x_5 &= 5 \end{aligned} \tag{7.17}$$

and the objective function

$$q = x_2 - x_1 \tag{7.18}$$

which seeks to find nonnegative values of the decision variables x_i minimizing (7.18) subject to (7.17).

It is often convenient to seek for LP solution by maximizing its objective function, what is an easy matter to do in this case by changing the sign

$$q' = -q = x_1 - x_2 \tag{7.19}$$

In this example, the number of equations is $m = 3$ and the number of unknowns is $n = 5$, so that there are $m = 3$ basic variables and $n - m = 2$ nonbasic variables. The fact that there are only two nonbasic variables allows us to illustrate the problem solution in two-dimensional space, i.e. on the plane.

The system (7.17) can be resolved for three variables, say x_3, x_4, and x_5 by expressing them in terms of x_1 and x_2, viz.,

$$\begin{aligned} x_3 &= 2 + 2x_1 - x_2 \\ x_4 &= 2 - x_1 + 2x_2 \\ x_5 &= 5 - x_1 - x_2 \end{aligned} \tag{7.20}$$

Recalling that the variables of an LP problem may take only nonnegative values, we obtain for the feasible solution region the constraining inequalities

$$x_i \geqslant 0, \quad i = 1, \ldots, 5 \tag{7.21}$$

Each of these inequalities defines some half-plane in the x_1, x_2 plane. The inequality $x_1 \geqslant 0$ defines the upper half-plane, and the inequali-

ty $x_3 \geqslant 0$ defines the half-plane lying on one side of the straight line $2 + 2x_1 - x_2 = 0$, precisely the one containing the origin (which may readily be verified by substituting the (0, 0) point in the first line of (7.20)). The half-plane of $x_3 < 0$ that does not contribute to the region of feasible solutions is indicated in Fig. 7.1 by hachures.

Similar constructions for all x_i are presented in Fig. 7.2. The straight lines corresponding to $x_i = 0$, $i = 1, \ldots, 5$, are labelled

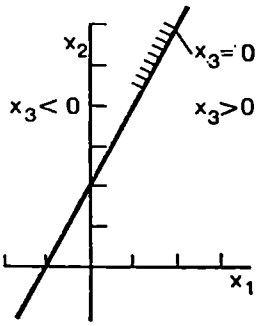


Fig. 7.1. A half-plane

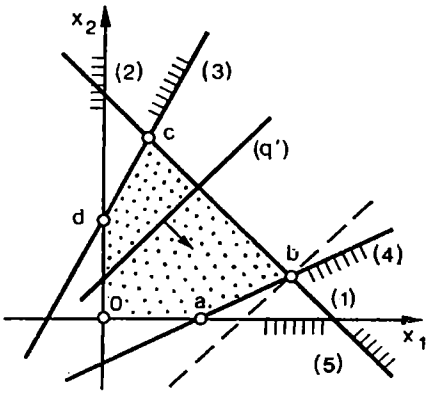


Fig. 7.2. Polygon of feasible solutions

a parenthesized digit. The areas of $x_i < 0$ are emphasized by hachures. The region of feasible solutions is the polygon Oabcd indicated by stippling. It is important to note that the polygon is closed and convex, because it represents an intersection of convex regions defined by conditions $x_i \geqslant 0$.

This construction leads us to a geometric interpretation of the basic solution. Since each line in Fig. 7.2 represents the locus of one variable being zero, then intersections of two lines are points where two, i.e. $n - m$, variables are zero. On the other hand, $n - m$ is the number of variables not in solution whose vanishing corresponds to basic solution. Hence, the intersections of the straight lines $x_i = 0$, $i = 1, \ldots, 5$, define the basic solutions of the LP problem.

Among the basic solutions there are some that do not belong to the region of feasible solutions. Only those points of intersection of the straight lines $x_i = 0$ belong to the region of feasible solutions that are the vertices of the polygon. Consequently, the vertices of the polygon of feasible solutions correspond to the feasible basic solutions.

Consider now conditions to maximize the objective function (7.19). Geometrically, for any q', equation (7.19) defines a straight line at 45° with the axis of abscissa, greater q' correspond to this line moving in the arrow indicated in Fig. 7.2. This line will be consistent with a feasible solution of the problem only if it has common points

with the region of feasible solutions. The value of q' will be a maximum in the extreme position of the line where it becomes a supporting line for the region of feasible solutions (hachured line in Fig. 7.2). However, a supporting line has to pass through at least one of vertices of the polygon, and the vertices, as we have seen, correspond to feasible basic solutions.

We have thus arrived at a very important conclusion that a solution of an LP problem, i.e. the one maximizing q', must be found among the *basic feasible solutions*. This conclusion may easily be generalized over the case of arbitrary m and n with $n - m > 2$. Now, equations (7.20) define hyperplanes in $(n - m)$-dimensional space, and the conditions (7.21) define half-spaces restricted by the hyperplanes. The feasible region will be a closed convex polyhedron. The set is *closed*, because the conditions (7.21) are specified as nonstrict inequalities, and *convex* because it is an intersection of convex sets (half-spaces).

The objective function (7.2) is linear, therefore it may be deemed both convex and concave. Hence, an optimum solution of LP is always unique and therefore *global*. The function q (x) has no stationary points. Consequently, the optimum solution is attained at the boundary of the feasible set.

Among boundary points of the feasible set there are *extreme points* which are the intersections of faces, i.e. represent basic solutions. A distinctive feature of extreme points is that they cannot be represented as a convex combination of other points of the set (see Sect. 4.3.3).

Let us assume that $\mathbf{cx}$ reaches a maximum at a boundary point $\mathbf{x}$ that is not an extreme point. We represent this point as a convex combination of a minimum number of extreme points so that this representation does not include extreme points of zero weight. Let these be points $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Then,

$$\mathbf{x} = \sum_i w_i \mathbf{x}_i, \quad w_i > 0, \quad \sum_i w_i = 1$$

and

$$\mathbf{cx} = \mathbf{c} \sum_i w_i \mathbf{x}_i = \sum_i w_i (\mathbf{cx}_i)$$

Assume that point $\mathbf{x}_k$ gives the greatest product $\mathbf{cx}_k$ of all extreme points. Denote this value of $\mathbf{x}$ by v. The right hand side of the expression for $\mathbf{cx}$ will not be smaller if we replace all $\mathbf{cx}_i$ by v. Consequently, $\mathbf{cx} \leqslant v$.

If $\mathbf{cx} < v$, then the optimal point is $\mathbf{x}_k$ rather than $\mathbf{x}$. The case of $\mathbf{cx} = v$ occurs only when the optimal solution is attained at all extreme points under consideration. Therefore, an optimal solution of an LP problem is attainable either in one extreme point, or on a set of extreme points, and hence on any convex combination of

these points. This implies that an LP solution must be among the basic feasible solutions.

This conclusion allows us to set up ways for LP problem solution. The solution must be among basic feasible solutions which are finite in number, therefore we can find all basic feasible solutions and calculate q' for each of them. The final solution will be the one of maximum q'. This way of solving an LP problem although feasible is rather laborious since the number of basic feasible solutions can be very large. There exist, however, more economical methods of exhaustive search over basic solutions which allow the analyst to deal with a minimum number of solutions. One of the most popular solution techniques of this type is the simplex method which will be considered in the coming section.

7.2 Solution of LP Problems

7.2.1 Simplex Computation

The simplex method of linear programming uses the basic concepts of matrix algebra to solve for the intersection of two or more lines or planes. It begins at some feasible solution, i.e. one that satisfies all constraints, and successively obtains solutions at intersections which offer better values of the objective function. The basic solution can be found by assuming some $n - m$ variables to be not in solution, equating them to zero and solving the system of equations (7.1). If some of the basic variables are negative, some other nonbasic variables should be chosen, i.e. a new basis should be adopted.

Once a feasible solution has been obtained, we check whether or not a maximum of the objective function q' is attained. If it is not the case, then we seek another feasible solution and such that increases q'. Repeating this procedure we see that it leads to the final solution in a rather fast manner that rejects a large number of basic solutions whose objective function is known to be far from a maximum. We illustrate this method with the example considered in subsection 7.13.

Consider the system of equations (7.17) which seeks for a nonnegative solution that maximizes the objective function (7.19). Since $n - m = 2$, we may take as nonbasic variables any two variables, say x_1 and x_2. Setting them equal to zero, the system (7.17) yields a basic solution $x_1 = x_2 = 0$, $x_3 = 2$, $x_4 = 2$, $x_5 = 5$, which is feasible and gives $q' = x_1 - x_2 = 0$.

We can check whether or not this solution maximizes the objective function by seeking another feasible solution giving a larger value of this function. In order to obtain a new feasible basic solution one of the variables, x_1 or x_2, should be made basic. It will then be nonzero, i.e. will increase in value. Hence, if any of the nonbasic vari-

ables enters in the expression for the objective function with the plus sign, i.e. contributes to the value of the objective function, then the maximum of the function has not been achieved and this variable should be made basic, i.e. other than zero.

When one nonbasic variable increases, some of basic variables decrease. Since nonnegative values of decision variables are inadmissible, a new nonbasic variable should be that basic variable which vanishes earlier than the others.

In our example, the objective function (7.19) includes a nonbasic variable x_1 with a plus sign. Hence, the objective function is not a maximum, and this variable should be made basic. To determine the new nonbasic variable we express the basic variables x_3, x_4, and x_5 in terms of the nonbasic variables by reducing (7.17) to (7.20). We see that at $x_2 = 0$ an increase in x_1 does not lead to a decrease in x_3, but decreases x_4 and x_5, so that at $x_1 = 2$ we have $x_4 = 0$, and $x_5 = 3$. Thus, x_4 should be made a nonbasic variable which leads to a new basis x_1, x_3, x_5.

Proceeding with the solution, we resolve the system (7.17) for the new variables

$$\begin{aligned} x_1 &= 2 + 2x_2 - x_4 \\ x_3 &= 6 + 3x_2 - 2x_4 \\ x_5 &= 3 - 3x_2 + x_4 \end{aligned} \tag{7.22}$$

The objective function will also be expressed through the new nonbasic variables x_2 and x_4,

$$q' = 2 + x_2 - x_4 \tag{7.23}$$

Repeating the earlier argument, we arrive at a conclusion that a maximum of the objective function has not been achieved and the nonbasic variable x_2 should be made basic, while the basic variable x_5 should be made nonbasic. The new basis will be x_1, x_2, x_3.

Resolving equation (7.17) for the new basic variables, we get

$$\begin{aligned} x_1 &= 4 - x_4/3 - 2x_5/3 \\ x_2 &= 1 + x_4/3 - x_5/3 \\ x_3 &= 9 - x_4 - x_5 \end{aligned} \tag{7.24}$$

Now the objective function takes the form

$$q' = 3 - 2x_4/3 - x_5/3 \tag{7.25}$$

that suggests that q' decreases in value as the nonbasic variables x_4 and x_5 increase. Consequently, a maximum of q' is attainable with this basis, and the solution of this LP problem will be $x_4 = x_5 = 0$, $x_1 = 4$, $x_2 = 1$, and $x_3 = 9$. These values give $q' = -q = 3$.

This method of LP problem solving has a disadvantage of unwieldy transformations of the system of linear equations. These transformations may be simplified by arraying these equations into tables

(matrices) of coefficients. A system transformation is then effected by recalculating the coefficients of these matrices. These calculations are carried out by formal rules that lend themselves well to computer coding.

7.2.2 Tabular Method of Optimization

Denote by x_i', $i = 1, \ldots, m$, basic variables, and by x_j'', $j = 1, \ldots, n-m$, nonbasic variables. If we rewrite the objective function and basic variables in terms of nonbasic variables, the LP problem formulation becomes as follows:

maximize

$$q' = -q = \alpha_{00} - \sum_{j=1}^{n-m} \alpha_{0j} x_j'' \tag{7.26}$$

subject to

$$x_i' = \alpha_{i0} - \sum_{j=1}^{n-m} \alpha_{ij} x_j'', \quad i = 1, \ldots, m, \quad x_i' \geqslant 0, \quad x_j'' \geqslant 0 \tag{7.27}$$

With this formulation the problem can be represented by the matrix of coefficients

$$\begin{array}{c} \\ q' \\ x_1' \\ \vdots \\ x_m' \end{array} \begin{array}{c} \begin{array}{cccc} 1 & -x_1'' & \ldots & -x_{n-m}'' \end{array} \\ \begin{pmatrix} \alpha_{00} & \alpha_{01} & \ldots & \alpha_{0,\,n-m} \\ \alpha_{10} & \alpha_{11} & \ldots & \alpha_{1,\,n-m} \\ \vdots & \vdots & & \vdots \\ \alpha_{m0} & \alpha_{m1} & \ldots & \alpha_{m,\,n-m} \end{pmatrix} \end{array} \tag{7.28}$$

This arrangement of the coefficients is convenient to judge whethre a basic solution is feasible and if so whether or not it is optimal. Indeed, noting that the column α_{i0}, $i \neq 0$, is a basic solution corresponding to the basis $x_1', \ldots, x_m'$ and the row of α_{0j}, $j \neq 0$, are the coefficients of nonbasic variables of q' taken with the opposite signs, we conclude that a basic solution corresponding to the basis $x_1', \ldots, x_m'$ is feasible if $\alpha_{i0} \geqslant 0$, $i \neq 0$. If in addition $\alpha_{0j} \geqslant 0$, $j \neq 0$, then this solution is also optimal. It is clear that for an optimal solution the coefficient α_{00} yields $q' = q_{\max} = -q_{\min}$.

To illustrate the procedure we begin with the evaluation of a feasible basic solution which in general is not optimal. Represent this solution in the form of a matrix in which element (ij) is the coefficient in the row of basic variable x_i and column of nonbasic variable x_j. To obtain a better basic solution this matrix of coefficients should be transformed. The transformation technique will be formulated as a system of rules illustrated by way of example solved in the preceding section. We avoid a rigorous definition of these rules

noting only that they correspond exactly to the transformation of the system of equations carried out for this example and may be verified by comparing the respective transformations.

Let the LP problem be given as the system of equations (7.17) and the objective function (7.19) to be maximized. Let x_1 and x_2 be nonbasic variables, then these equations and the objective function reduce to

$$\begin{aligned} q' &= 0 - (-x_1 + x_2) \\ x_3 &= 2 - (-2x_1 + x_2) \\ x_4 &= 2 - (x_1 - 2x_2) \\ x_5 &= 5 - (x_1 + x_2) \end{aligned} \tag{7.29}$$

The matrix of coefficients is arrayed in form (a) of Table 7.2, where the upper corner of each cell contains coefficients α_{ij} of Eqs. (7.29).

Table 7.2

(a)

	1	$-x_1$	$-x_2$
q'	0 / 2	−1 / 1	1 / −2
x_3	2 / 4	−2 / 2	1 / −4
x_4	2 / 2	① / 1	−2 / −2
x_5	5 / −2	1 / −1	1 / 2

(b)

	1	$-x_4$	$-x_2$
q'	2 / 1	1 / $-\frac{1}{3}$	−1 / $\frac{1}{3}$
x_3	6 / 3	2 / −1	−3 / 1
x_1	2 / 2	1 / $-\frac{2}{3}$	−2 / $\frac{2}{3}$
x_5	3 / 1	−1 / $-\frac{1}{3}$	③ / $\frac{1}{3}$

(c)

	1	$-x_4$	$-x_5$
q'	3	$\frac{2}{3}$	$\frac{1}{3}$
x_3	9	1	1
x_1	4	$\frac{1}{3}$	$\frac{2}{3}$
x_2	1	$-\frac{1}{3}$	$\frac{1}{3}$

Let us see whether we have found an optimal solution satisfying $\alpha_{0j} \geqslant 0$, $j \neq 0$. Since α_{01}, the coefficient of $-x_1$ in q', is nonnegative, we conclude that the optimal solution is not obtained and the variable x_1 should be made basic. If the coefficients α_{0j} of several nonbasic variables turn out to be negative, then any of them may be converted into a basic variable.

Determine now which of the basic variables should be made nonbasic. Clearly, this must be one that vanishes first in increasing x_1. In other words this will be the basic variable x_i for which the coefficient $\alpha_{i1} > 0$ and the ratio α_{i0}/α_{i1} is least of all. In our case this variable is x_4 with $\alpha_{40} = 2$ and $\alpha_{41} = 1$.

The circled coefficient α_{41} in the upper corner of the cell in the emphasized row and column is called the *pivot*. For convenience denote $1/\alpha_{41} = \lambda$; in our example $\lambda = 1$.

Now, we fill bottom corners of the cells according to the replacement rule as follows. Divide the pivotal row by the pivot and write the results in bottom corners (upper coefficients except the pivot are printed in boldface). The cells of the pivotal column receive the upper coefficients multiplied by $-\lambda$ (bottom coefficients, except the pivot, are printed in boldface). In other cells we write the product of boldfaced coefficients whose intersection forms the particular cell.

Table 7.2b, the next step of the procedure, differs from table (a) in that the nonbasic variable x_1 becomes basic and the basic variable x_4 becomes nonbasic. Upper corners are filled as follows. Write in the row and column corresponding to the new nonbasic and basic variables the bottom coefficients of the pivotal row and column of table (a). The other cells are filled with the sums of the coefficients in the respective cells of table (a). Table (b) filled in this way corresponds to the matrix of coefficients (7.28) in the new basis x_1, x_3, x_5.

Because $\alpha_{02} < 0$ in table (b) no optimal solution has been found. Consequently, the procedure should be continued filling bottom corners of table (b) to obtain a new table (c) corresponding to the basis x_1, x_2, x_3. In this table the coefficients a_{0j}, $j \neq 0$, are positive, and it yields an optimal solution that can be evaluated from the column of the nonbasic terms, namely, $x_1 = 4$, $x_2 = 1$, $x_3 = 9$, $x_4 = x_5 = 0$, $q'_{\max} = -q_{\min} = 3$.

7.2.3 Determination of Basic Feasible Solutions

In LP problems with a large number of equations and unknowns the evaluation of an initial basic feasible solution is a toll task. One of solutions to this problem is given below [38].

The general form of a system of equations having a basic feasible solution $x'_1, \ldots, x'_m$ can be obtained from (7.27) in the form

$$x'_i + \sum_{j=1}^{n-m} \alpha_{ij} x''_j = \alpha_{i0}, \quad \alpha_{i0} \geqslant 0, \quad i = 1, \ldots, m \tag{7.30}$$

In this system of equations, each variable x'_i taken as basic enters only one equation with coefficient $+1$ (or with any positive coefficient as both sides of this equation may be divided by this coefficient). If such variables can be found in each equation of (7.1), then they set up a starting feasible basis. In the above example with the system (7.17) these variables were x_3, x_4, and x_5.

If some of the equations have no such variables, then the procedure may be as follows. Renumber the variables that may be taken as

basic, viz. $x'_1, \ldots, x'_s$, $s < m$. In the remaining $m - s$ equations we introduce artificial basic variables $\hat{x}'_k$, $k = s + 1, \ldots, m$, assuming they are nonnegative, $\hat{x}'_k \geqslant 0$. As the result of this transformation the system of equations takes the form

$$\begin{aligned} x'_i + \sum_{j=1}^{n-s} \alpha_{ij} x''_j &= \alpha_{i0}, \quad i = 1, \ldots, s \\ \hat{x}'_k + \sum_{j=1}^{n-s} \alpha_{kj} x''_j &= \alpha_{k0}, \quad k = s+1, \ldots, m \end{aligned} \tag{7.31}$$

In order for this system to coincide with the initial one, the artificial variables must vanish in the final solution. For this purpose they receive sufficiently large negative coefficients in the expression for q'.

7.2.4 Dual Program

Let us revisit the problem of limited resource allocation considered in Example 7.1 and look from the standpoint of an enterprise at the worth of resources. Clearly, resources that cannot be used completely are of extremely low value in the sense that the enterprise will be reluctant to pay for an increase of these resources. For example, costly equipment not involved in the production process is of zero value for the manufacturer. Most valuable resources will be those governing production most of all, i.e. controlling the profit, so that the manufacturer is ready for considerable expenses to increase the stock of such resources.

Accordingly we may assume that each type of resource possesses some marginal or "shadow" price [38] that defines its value for the activity from the standpoint of profit. This value also depends on available stock of the particular resource and demand of this resource in the production process.

If the enterprise has to limit itself to one technological process that requires large amounts of some resource of limited stock, then the shadow price of such resource will be large. However, shadow prices set up for this process will not be best as introduction of other technological processes will lead to more rational usage of all stocks. Hence, there exist optimal shadow prices that correspond to some maximum profit of the enterprise, i.e., optimal allocation of resources. Reflecting this fact, the economics literature refers to optimal shadow prices as *optimal estimates*. The evaluation of optimal shadow prices is thus closely connected with the problem of optimal allocation of resources, i.e. with the LP problem defined by the system of equations (7.5) and the objective function (7.6). Optimal sha-

dow prices may be determined in an LP problem of different formulation.

Denote by u_i the shadow price of unit of resource i. Values of u_i should be such that the shadow price of resources used in any production process could not be less than the respective profit. With the nomenclature of Example 7.1 and data of Table 7.1 this condition may be written as

$$\sum_{i=1}^{m} a_{ij}u_i \geqslant c_j, \quad j=1, \ldots, l \tag{7.32}$$

If we introduce variables $u_{m+j} \geqslant 0$ representing the excess of unit price of a product over the profit from its realization, then the system of inequalities (7.32) becomes

$$\sum_{i=1}^{m} a_{ij}u_i - u_{m+j} = c_j, \quad j=1, \ldots, l \tag{7.33}$$

Optimal shadow prices will be such that minimize the total cost of resources, i.e. the quantity

$$q^* = \sum_{i=1}^{m} b_i u_i \tag{7.34}$$

The constraints (7.32) along with the objective function (7.34) make up a new problem of linear programming that is called the *dual program* to the problem of Example 7.1 called the *primal program*.

It is not hard to see that the primal and dual problems are intimately connected. If the primal program is a maximization problem, then its dual is a minimization problem. The coefficients of the objective function in the primal program are nonbasic terms in the constraints of the dual; the nonbasic terms of the constraints of the primal program become the coefficients of the objective function in the dual. The coefficients of variables in the constraints of the dual program become the columns in the table of the primal problem. The signs of the inequalities change for the opposite.

A close relationship exists also between the solutions of primal and dual problems. In order to establish this relationship we write the equations of the primal program as

maximize $$'q' = 0 - \sum_{j=1}^{l} (-c_j)\, x_j \tag{7.35}$$

subject to $$x_{l+i} = b_i - \sum_{j=1}^{l} a_{ij}x_j, \quad i=1, \ldots, m \tag{7.36}$$

where $x_1, \ldots, x_l$ are nonbasic variables.

The associated matrix is

$$\begin{array}{c} \\ q' \\ x_{l+1} \\ \vdots \\ x_{l+m} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & -x_1 & \dots & x_l \end{array} \\ \begin{pmatrix} 0 & -c_1 & \dots & -c_l \\ b_1 & a_{11} & \dots & a_{1l} \\ \vdots & \vdots & & \vdots \\ b_m & a_{m1} & \dots & a_{ml} \end{pmatrix} \end{array} \tag{7.37}$$

In the dual program we take $u_1, \dots, u_m$ as nonbasic variables so that its formulation becomes

$$\text{minimize} \qquad q^* = 0 + \sum_{i=1}^{m} b_i u_i \tag{7.38}$$

$$\text{subject to} \qquad u_{m+j} = -c_j + \sum_{i=1}^{m} a_{ij} u_i, \quad j = 1, \dots, l \tag{7.39}$$

The associated matrix is

$$\begin{array}{c} \\ q^* \\ u_{m+1} \\ \vdots \\ u_{m+l} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & u_1 & \dots & u_m \end{array} \\ \begin{pmatrix} 0 & b_1 & \dots & b_m \\ -c_1 & a_{11} & \dots & a_{1m} \\ \vdots & \vdots & & \vdots \\ -c_l & a_{l1} & \dots & a_{ml} \end{pmatrix} \end{array} \tag{7.40}$$

As can be seen, the columns of matrix (7.40) are the rows of (7.37). Hence, both primal and dual programs may be described by one and the same matrix (7.37), in which the correspondence between the primal and dual variables is as follows:

$$\begin{aligned} x_j &\leftrightarrow -u_{m+j}, \quad j = 1, \dots, l \\ x_{l+i} &\leftrightarrow u_i, \quad i = 1, \dots, m \end{aligned} \tag{7.41}$$

It is worth noting that any transformation of matrix (7.37) under the above rules will lead to a new matrix, which also will describe both primal and dual programs. Hence, the general form of a matrix (7.28) may be used to describe both primal and dual LP problems. If the elements of the first column (except possibly α_{00}) are positive, then the matrix corresponds to a basic feasible solution of the primal problem. If the elements of the first row are positive (except possibly α_{00}), then the matrix corresponds to a feasible solution of the dual problem. If the elements in the first column and upper row of matrix (7.28) are positive (least possibly α_{00}), then the matrix corresponds to the optimal solution of both primal and dual programs. The coefficient α_{00} gives the value of the objective function, that is the same for the optimal solution of primal and dual programs,

$$q'_{\max} = q^*_{\min} \tag{7.42}$$

Example 7.5. The resource allocation problem is given by Table 7.3. The table also shows the solution of the primal and dual

Table 7.3

Resources	kg/piece			In stock	Marginal	Shadow prices
	A	*B*	*C*			
R_1	2	1	1	25	5.5	0
R_2	1	1	1	14	0	3.0
R_3	0	4	2	19	0	0.5
R_4	3	0	1	24	0	1.0
Profit per piece	6	5	5	—	—	—
Optimal allocation	5.5	1.0	7.5	—	—	—

programs, i.e. the optimal allocation of resources, marginal resources at the optimal allocation and marginal (shadow) prices. It is seen that excessive type R_1 resource has sero price, and type R_2 resource involved in all production processes and available in low stock only has the highest price.

We note in conclusion that shadow prices may in some situations be a valuable tool of management. For example, in a large enterprise or industry where many decisions are relegated to departments or shops, it should be difficult to inform each department about decisions made by other departments or shops. In such situations shadow prices may provide each department with a valuable estimate for making decisions close to optimal from the standpoint of the entire enterprise or industry.

7.2.5 The Notion of Integer Programming

In many LPPs variables are integer quantities. For example, an activity cannot produce 7.5 tools or 4.8 cars. In this type of problem the conditions (7.1) and (7.2) are augmented by the requirement that variables x_i be integer quantities, or be elements of a finite set in Example 7.4. Geometrically this implies that feasible solutions will occupy not the entire feasible region defined by the constraints (7.1) but only discrete points of this region as illustrated in Fig. 7.3.

One can, of course, try to solve such a problem overlooking the condition that the variables be integers and find a solution defined

by, say, point *1*, then round off this value to the nearest integer to get point *2* as an IP solution. However, this solution might be far from the optimal one. The optimal solution in this example may be point *3*.

Amidst the algorithms available for IP problem solving Gomory's method is most popular. We do not intend to dwell here on the computational possibilities of this method, referring the reader to the books of Kuzik [5] and Saaty [41], and confine ourselves to outlining the idea of the method.

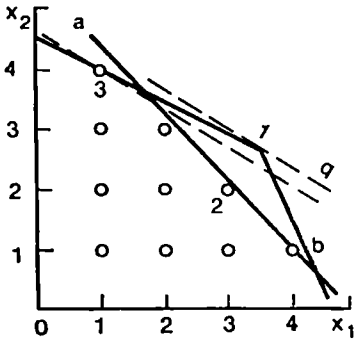


Fig. 7.3. Feasible solution set in an IP problem

The Gomory approach is based on the simplex method that performs optimization disregarding the integer requirement of the variables. If the solution is a noninteger, then an additional constraint is introduced which in two dimensions is a straight line (*ab* in Fig. 7.3) or a cutting hyperplane in the general case. This cutting hyperplane cuts off an area of the feasible region along with the optimal solution, which does not contain integer-valued feasible points. With this additional constraint, the found optimal solution will be outside the new feasible region, therefore the problem is again solved by the simplex method and an optimal solution is sought in the new feasible region. If this solution is again a noninteger, the procedure is repeated.

Recent developments for IP problem solution are combinatorial methods, notable the branch and bound algorithm [41].

7.3 Nonlinear Programming

7.3.1 Problem Formulation

The nonlinear programming (NLP) problem seeks to find a value of multidimensional variable $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})$ which minimizes the objective function $q(\mathbf{x})$ subject to

$$f_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, m \tag{7.43}$$

and

$$x^{(j)} \geqslant 0 \tag{7.44}$$

Some formulations have opposite signs in the inequalities (7.43). However, recognizing that the opposite inequality may be obtained by multiplying by -1, a problem may be always reduced to one ha-

ving constraints of the same sign. If some of the constraints are equations, say $\varphi(\mathbf{x}) = 0$, then they may be represented by the pair of inequalities $\varphi(x) \leqslant 0$ and $-\varphi(x) \leqslant 0$, retaining the typical form of the model.

7.3.2 Penalty Function Algorithms

As noted in Sect. 6.6, the constrained optimization problems with equality constraints may be reduced to unconstrained problems by modifying the objective function with the aid of Lagrange multipliers. In a very similar manner, the problem of minimization of an objective function $q(\mathbf{x})$ subject to inequality constraints (7.43) may be reduced to an unconstrained optimization problem by modifying the objective function with addends in the form of appropriately selected penalty functions [36].

The general idea of the penalty function algorithm consists in constructing a sequence of new objective functions

$$Q_k(\mathbf{x}) = q(\mathbf{x}) + r_k \varphi(\mathbf{x}), \qquad k = 1, 2, \ldots$$

where $\varphi(\mathbf{x})$ is the penalty function that assumes possibly small (better zero) values within the feasible region and increases beyond this region, thus penalizing constraint violation, and r_k, $k = 1, 2$, etc., is an increasing sequence of positive numbers, penalty parameters.

The most popular type of penalty function under constraints (7.43) is the expression

$$\varphi(\mathbf{x}) = \sum_{i=1}^{m} (f_i^+(\mathbf{x}))^2$$

where $f_i^+(\mathbf{x})$ denotes $f_i(\mathbf{x})$ if $f_i(\mathbf{x}) > 0$, and 0 if $f_i(\mathbf{x}) \leqslant 0$, i.e.

$$f_i^+(\mathbf{x}) = \max [0, f_i(\mathbf{x})]$$

The penalty function algorithm may be summarized as follows. Choose an arbitrary initial approximation $\mathbf{x}_0$ and a monotonously increasing sequence of real numbers r_k that tends to infinity. For $k = 1, 2$, etc. solve unconstrained minimization problem beginning with $\mathbf{x}_{k-1}$ for $Q_k(\mathbf{x})$ to find an approximation $\mathbf{x}_k$ to the solution of the original problem.

Convergence of this algorithm and other types of penalty function are discussed in depth by Moiseyev *et al* [36].

Example 7.6. Let x be a scalar and $q(x) = x^2 - 4(x - 1)$ as plotted in Fig. 7.4. Find x^* minimizing $q(x)$ subject to $x - 1 \leqslant 0$. The feasible region is shown in Fig. 7.4 by hachures. An obvious solution is $x^* = 1$.

The modified objective function is

$$Q_k(x) = x^2 - 4(x - 1) + r_k(\max(0, x - 1))^2$$

The derivative of this function is

$$Q_k'(x) = 2x - 4 \qquad \text{for } x \leqslant 1$$
$$= 2x - 4 + 2r_k(x - 1) \quad \text{for } x > 1$$

It is an easy matter to see that $Q_k'(x)$ vanishes in a single point $x_k = (2 + r_k)/(1 + r_k)$ that lies outside the feasible region of $x \leqslant 1$. However, as $r_k \to \infty$, x_k approaches $x^* = 1$.

When constraint qualifications are more involved than in Example 7.6 and the variable is multidimensional, evaluation of an initial approximation and optimal solutions for modified objective functions $Q_k(\mathbf{x})$ encounters a number of difficulties. Therefore, it is very important to have necessary conditions on the point optimizing the objective function subject to all constraints. Although such conditions are normally not sufficient, they facilitate the search for optimal solution.

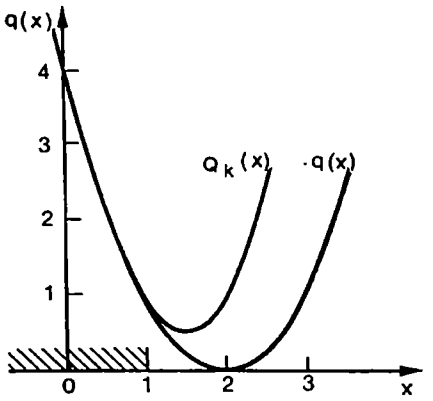


Fig. 7.4. Objective function $q(x)$ and the modified objective function $Q_k(x)$

In the classical unconstrained optimization problem, the necessary conditions for an optimal solution are those of (6.71), i.e. conditions defining stationary points of the objective function $q(\mathbf{x})$. For the NLP problem, such conditions are those of Kuhn-Tucker. Before we come to formulating these conditions we consider a simpler version of NLP problem.

7.3.3 Equality Constraints and Nonnegativity of Variables

The simplest NLP problem is minimization of $q(\mathbf{x})$ subject to the equality constraints.

$$f_j(\mathbf{x}) = 0, \quad j = 1, \ldots, m \tag{7.45}$$

and

$$x^{(i)} \geqslant 0, \quad i = 1, \ldots, n$$

or

$$\mathbf{x} \geqslant 0 \tag{7.46}$$

This problem may be viewed as an expansion of the conditioned minimization problem for nonnegative variables considered in

Section 6.6, so that the former may be analyzed with the familiar formalism of the Lagrangian function $L(\mathbf{x}, \lambda)$ defined by the relationship (6.76).

Because this problem formulation includes inequality constraints the relationships (6.77) hold here as well, namely, at the point $\mathbf{x}$ of optimal solution

$$L(\mathbf{x}, \lambda) = q(\mathbf{x}) \tag{7.47}$$

Accordingly, analysis of the behaviour of the Lagrangian function in the neighborhood of an optimal point may be replaced with the investigation of the objective function $q(\mathbf{x})$.

Let $\mathbf{x}$ be a point corresponding to an optimal solution. It may be either internal or boundary point of the feasible region $\mathbf{x} \geqslant 0$, i.e. each its component, say $x^{(i)}$, will satisfy either $x^{(i)} > 0$ or $x^{(i)} = 0$.

If $x^{(i)}$ does not lie on the boundary of the feasible region, i.e. $x^{(i)} > 0$, then deviations from $\mathbf{x}$ may be toward either increasing or decreasing $x^{(i)}$. However, since $\mathbf{x}$ is an optimum point, then there must be $\partial q(\mathbf{x})/\partial x^{(i)} = 0$.

If $x^{(i)}$ lies on the boundary of the feasible region, i.e. $x^{(i)} = 0$, then deviations from the optimal point may be only toward increasing $x^{(i)}$, and $q(\mathbf{x})$ must be increasing so that $\partial q(\mathbf{x})/\partial x^{(i)} > 0$.

If we generalize these situations observing (7.47) and (6.78), then necessary conditions that $\mathbf{x}$ is a solution of this optimization problem may be written as

$$\begin{aligned} \frac{\partial L(\mathbf{x}, \lambda)}{\partial x^{(i)}} &= 0, \quad \text{if } x^{(i)} > 0 \\ &> 0, \quad \text{if } x^{(i)} = 0, \quad i = 1, \ldots, n \end{aligned} \tag{7.48}$$

$$\partial L(\mathbf{x}, \lambda)/\partial \lambda_j = 0, \quad j = 1, \ldots, m \tag{7.49}$$

7.3.4 Kuhn-Tucker Conditions

Here we shall formulate necessary conditions that an optimal solution of the NLP problem is achieved at point $\mathbf{x}$, i.e. this point minimizes $q(\mathbf{x})$ subject to the constraints (7.43) and (7.44). To begin with we simplify the problem by introducing slack variables z_j in the inequality constraints to make them equality constraints,

$$f_j(\mathbf{x}) + z_j = 0, \quad z_j \geqslant 0, \quad x^{(i)} \geqslant 0 \tag{7.50}$$

with

$$\begin{aligned} z_j &> 0, \quad \text{if } f_j(\mathbf{x}) < 0 \\ &= 0, \quad \text{if } f_j(\mathbf{x}) = 0 \end{aligned} \tag{7.51}$$

The problem is seen to reduce to the considered model with equality constraints with $n + m$ variables z_j, $x^{(i)}$, $j = 1, \ldots, m$, $i =$

$1, \ldots, n$, and the Lagrangian function

$$L'(\mathbf{x}, \mathbf{z}, \lambda) = q(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j [f_j(\mathbf{x}) + z_j] \tag{7.52}$$

Conditions that must be satisfied by z_j and $x^{(i)}$ at the point of optimal solution will be expressed by the relationships (7.48) and (7.49) imposed on $L(\mathbf{x}, \mathbf{z}, \lambda)$.

The condition (7.48) imposed on z_j gives

$$\lambda_j = \frac{\partial L(\mathbf{x}, \mathbf{z}, \lambda)}{\partial z_j} = 0, \quad \text{if} \quad z_j > 0$$
$$> 0, \quad \text{if} \quad z_j = 0 \tag{7.53}$$

This relationship imposes constraints on the Lagrange multipliers λ_j. It implies that inequality constraints will not enter $L(\mathbf{x}, \mathbf{z}, \lambda)$ since for them $\lambda_j = 0$, and equality constraints corresponding to $z_j = 0$ will enter it with $\lambda_j > 0$. Therefore $L(\mathbf{x}, \mathbf{z}, \lambda)$ practically coincides with $L(\mathbf{x}, \lambda)$ defined by (6.76). The optimality condition (7.48) for $x^{(i)}$ retains its form.

If we take into account that $\partial L(\mathbf{x}, \lambda)/\partial \lambda_i = f_j(\mathbf{x})$ and the constraints on λ_j that follow from (7.53), then the final condition that must be satisfied by a point $\mathbf{x}$ being the solution of the general NLP problem may be written as

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial x^{(i)}} > 0, \quad \text{if} \quad x^{(i)} = 0$$
$$= 0, \quad \text{if} \quad x^{(i)} > 0, \quad i = 1, \ldots, n \tag{7.54}$$

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda_j} < 0, \quad \text{if} \quad \lambda_j = 0$$
$$= 0, \quad \text{if} \quad \lambda_j > 0, \quad j = 1, \ldots, m \tag{7.55}$$

These conditions are called the *Kuhn-Tucker conditions*. Sometimes they are written in another form. Define

$$\partial L(\mathbf{x}, \lambda)/\partial x^i = p_i$$
$$\partial L(\mathbf{x}, \lambda)/\partial \lambda_j = -q_j \tag{7.56}$$

Then (7.54) and (7.55) become

$$x^{(i)} \geqslant 0, \; p_i \geqslant 0, \; x^{(i)} p_i = 0, \; i = 1, \ldots, n$$
$$\lambda_j \geqslant 0, \; q_j \geqslant 0, \quad \lambda_j q_j = 0, \; j = 1, \ldots, m \tag{7.57}$$

Another interpretation of the Kuhn-Tucker conditions is associated with the saddle point of the Lagrangian function $L(\mathbf{x}, \lambda)$. The *saddle point* of $L(\mathbf{x}, \lambda)$ is a point where $L(\mathbf{x}, \lambda)$ is a minimum in $\mathbf{x}$ and a maximum in λ, i.e. a point $(\mathbf{x}_0, \lambda_0)$ satisfying

$$L(\mathbf{x}_0, \lambda) \leqslant L(\mathbf{x}_0, \lambda_0) \leqslant L(\mathbf{x}, \lambda_0) \tag{7.58}$$

Conditions that the saddle point of L (x, λ) must meet coincide with the Kuhn-Tucker conditions. Therefore, solution of the NLP problem may be treated as the evaluation of saddle points for the Lagrangian function.

It is instructive to note that in the general case the Kuhn-Tucker conditions are but the necessary conditions for a solution to the NLP problem to exist. In order that they could be sufficient it is necessary that the objective function have a single optimum in the feasible region and this optimal point will in this case be global. To satisfy this condition the objective function must be convex and the system of constraints should be such that the feasible region would be a close convex set. In addition, the objective function q (x) and functions f_i (x) defining the problem constraints must be continuously differentiable.

If the necessary and sufficient conditions of existence of an NLP problem solution are satisfied, then efficient computational techniques can be found for NLP problem solving. One of such techniques is outlined below.

7.3.5 Quadratic Programming

The quadratic program is an NLP problem which seeks to minimize the sum of linear and quadratic forms subject to constraints in the form of linear inequalities and nonnegativity conditions. Written in matrix form it is to

$$\begin{aligned} &\text{minimize } q(\mathbf{x}) = \mathbf{cx} + \mathbf{x}^{\mathrm{T}}\mathbf{Dx}, \\ &\text{subject to } \mathbf{Ax} \leqslant \mathbf{b} \\ &\text{and } \mathbf{x} \geqslant 0 \end{aligned} \tag{7.59}$$

It is assumed that the problem has m constraints and n variables, so that matrix **A** is of $m \times n$ type, **D** of $n \times n$ type, **b** is an m-vector, and **x** and **c** are n-vectors. In the expanded form these relationships are to

$$\begin{aligned} &\text{minimize} && q(\mathbf{x}) = \sum_j c_j x^{(j)} + \sum_j \sum_h d_{jh} x^{(i)} x^{(h)} \\ &\text{subject to} && \sum_h a_{jh} x^{(h)} - b_j \leqslant 0, \quad j = 1, \ldots, m \end{aligned} \tag{7.60}$$

and $x^{(i)} \geqslant 0$, $i = 1, \ldots, n$.

We limit our consideration to the case when the quadratic form is positive definite, i.e. convex. The linear form is also a convex function. Therefore, the objective function will be convex. In this case the necessary Kuhn-Tucker conditions will be also sufficient conditions for a single optimal point to exist.

To write the Kuhn-Tucker conditions we introduce the Lagrangian function. In view of (6.76) and (7.60)

$$L(\mathbf{x}, \lambda) = \sum_j c_j x^{(j)} + \sum_j \sum_h d_{jh} x^{(j)} x^{(h)} + \sum_j \lambda_j \left(\sum_h a_{jh} x^{(h)} - b_j \right) \qquad (7.61)$$

The derivatives of $L(\mathbf{x}, \lambda)$ with respect to $x^{(i)}$ and λ_j using the notation of (7.56) may be written as

$$\begin{aligned} c_i + 2 \sum_h d_{ih} x^{(h)} + \sum_j \lambda_j a_{ji} &= p_i, \quad i = 1, \ldots, n \\ b_j - \sum_h a_{jh} x^{(h)} &= q_j, \quad j = 1, \ldots, m \end{aligned} \qquad (7.62)$$

According to the Kuhn-Tucker conditions the solution of these equations should be sought subject to the requirements (7.57) which can be written in the form

$$\begin{gathered} x^{(i)} p_i = 0, \quad \lambda_j q_j = 0 \\ x^{(i)} \geqslant 0, \quad \lambda_j \geqslant 0, \quad p_i \geqslant 0 \\ q_j \geqslant 0, \quad i = 1, \ldots, n; \quad j = 1, \ldots, m \end{gathered} \qquad (7.63)$$

The system (7.62) has $n + m$ equations in $2(n + m)$ variables $x^{(i)}$, λ_j, p_i, q_i, of which $n + m$ are nonbasic and equal to zero. The other variables form a basic feasible solution if conditions (7.63) are satisfied.

If the number of variables in the problem is not large, then one may try to guess a basic feasible solution by letting arbitrarily $n + m$ variables to be nonbasic, i.e. equating them to zero, and by solving the system (7.62) find the values of the basic variables. This approach, however, fails to guarantee that the obtained values of variables will satisfy (7.63). Therefore, attempts to guess a basic feasible solution have to be multiply repeated.

If a basic feasible solution is found, then it can be improved, and a new better basis obtained by a version of the simplex method similar to that considered in linear programming. The difference is that when a new basic variable is selected one must verify that the conditions $x^{(i)} p_i = 0$ and $\lambda_j q_j = 0$ are satisfied. These conditions imply that if the basis has $x^{(i)}$ or λ_j, then it cannot incorporate p_i and q_j, respectively.

Where the number of variables is large, guessing a basic feasible solution may be extremely cumbersome. A way out may be the efficient systematic techniques of obtaining a basic feasible solution outlined in books by Hadley [39] and Yudin [47].

If the quadratic program formulation (7.59) has linear constraints given as equalities $\mathbf{Ax} = \mathbf{b}$, then the Kuhn-Tucker conditions will be written in the form (7.48) and (7.49). The condition (7.49) implies

$q_j = 0$, $j = 1, \ldots, m$, in (7.62) and (7.63). This considerably simplifies problem solution, as the application of the simplex method will then require only verification that $x^{(i)} p_i = 0$.

7.4 Iterative Optimization Techniques

7.4.1 Problem Formulation

The limited capabilities of the simplex method in application to problems with complex forms of constraints and arbitrary forms of objective function have motivated wide use of iterative methods of search for optimal solution based on the concept of gradient of objective function $q(\mathbf{x})$.

The *gradient* of a function $q(\mathbf{x})$, denoted by grad $q(\mathbf{x})$ or $\nabla q(\mathbf{x})$, is a vector whose magnitude represents the rate of variation of $q(\mathbf{x})$, and the direction coincides with one in which this function increases most of all. On the other hand, the vector $\nabla q(\mathbf{x})$ showing the direction in which $q(\mathbf{x})$ decreases most of all is sometimes called the *antigradient* of $q(\mathbf{x})$.

Let $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^{\mathrm{T}}$, then the gradient of $q(\mathbf{x})$ is the column-vector

$$\nabla q(\mathbf{x}) = \left(\frac{\partial q(\mathbf{x})}{\partial x^{(1)}}, \ldots, \frac{\partial q(\mathbf{x})}{\partial x^{(n)}} \right)^{\mathrm{T}} \tag{7.64}$$

The condition for a point $\mathbf{x}$ to be a stationary point of $q(\mathbf{x})$ may be written now as

$$\nabla q(\mathbf{x}) = 0 \tag{7.65}$$

As a first step towards iterative optimization algorithms we expand $q(\mathbf{x})$ in a Taylor series in the vicinity of the optimal point $\mathbf{x}^*$ which is deemed stationary. Using (4.107), we have

$$q(\mathbf{x}) = q(\mathbf{x}^*) + \frac{1}{2} \sum_h \sum_j a_{hj} \Delta x^{(h)} \Delta x^{(j)} \tag{7.66}$$

where $\Delta \mathbf{x} = \mathbf{x}^* - \mathbf{x}$ is the deviation from the optimal point, and

$$a_{hj} = a_{jh} = \frac{\partial^2 q(\mathbf{x})}{\partial x^{(h)} \partial x^{(j)}} \Big|_{\mathbf{x}=\mathbf{x}^*} \tag{7.67}$$

Taking the derivatives of $q(\mathbf{x})$ with respect to $x^{(i)}$, $i = 1, \ldots, n$, we find the component of the gradient of $q(\mathbf{x})$ at point $\mathbf{x}$ under consideration

$$\frac{\partial q(\mathbf{x})}{\partial x^{(i)}} = - \sum_h a_{hi} \Delta x^{(h)}, \quad i = 1, \ldots, n \tag{7.68}$$

The expression for the gradient of $q(x)$ is

$$\nabla q(x) = - \mathbf{A} \Delta \mathbf{x} = - \mathbf{A} (\mathbf{x}^* - \mathbf{x}) \tag{7.69}$$

where A is the $n \times n$ symmetric matrix with elements a_{hj} defined by (7.67). Resolving the system (7.69) for x*, we have

$$\mathbf{x}^* = \mathbf{x} - \mathbf{A}^{-1} \nabla q(\mathbf{x}) \tag{7.70}$$

If the expansion (7.66) were an exact representation of $q(\mathbf{x})$ and there existed simple methods to determine A, then the value of x* could be found directly from (7.70). In most practical problems, however, these conditions are not satisfied.

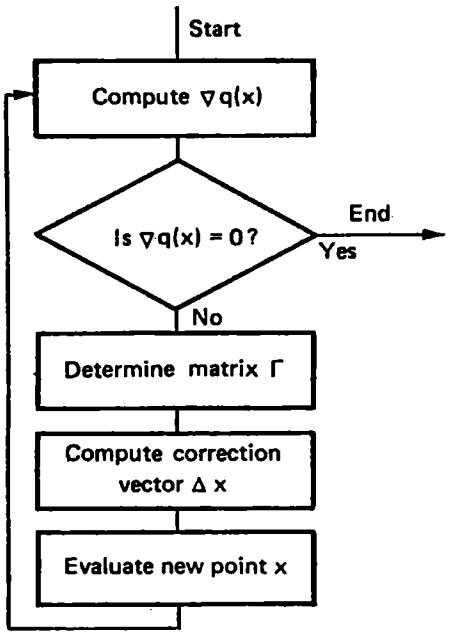


Fig. 7.5. Optimum search algorithm

If we substitute matrix Γ with elements γ_{hj} for the unknown matrix $\mathbf{A}^{-1}$, then we may expect that under an appropriate choice of coefficient γ_{hj} the quantity $\mathbf{x} - \Gamma \nabla q(\mathbf{x})$ will give a value closer to the optimal point than x does. This opens up a possibility for multi-step search procedure.

Denote by $\mathbf{x}_n$ the value of x at the nth step and assume that the element of matrix Γ may be selected differently at different steps of the procedure. Then in agreement with (7.70) we write the procedure of multi-step search for an optimum of $q(\mathbf{x})$ as

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \Gamma_n \nabla q(\mathbf{x}_n) \tag{7.71}$$

In order for the sequence of $\mathbf{x}_n$ to lead to an optimal solution, the coefficients of Γ_n must satisfy certain conditions of convergence. Clearly, various choices of Γ_n result in different optimization algorithms.

Sometimes the optimization algorithm is written in the following basic form

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \Delta \mathbf{x}_n \tag{7.72}$$

where the correction vector

$$\Delta \mathbf{x}_n = -\Gamma_n \nabla q(\mathbf{x}_n) \tag{7.73}$$

The algorithm built on the expressions (7.72) and (7.73) is shown in Fig. 7.5.

Graphical illustrations of the gradient methods discussed below will be provided for the case of $\mathbf{x} = (x^{(1)}, x^{(2)})$. In minimization problems, the objective function may be pictured as a surface for-

med into a valley, as shown in Fig. 7.6*a*, and called the response surface. A two-dimensional representation of a three-dimensional

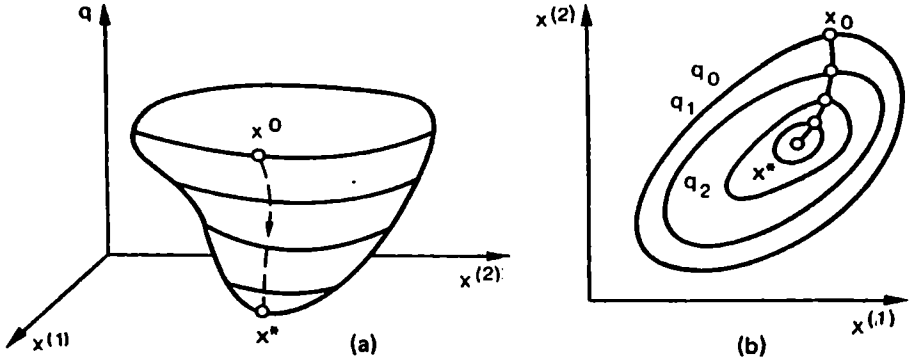


Fig. 7.6. (*a*) Response surface and (*b*) representation of this surface by level lines

response surface by q-level lines, being cross sections of the valley by horizontal planes, is shown in the diagram at (*b*). The respective value of q is shown behind each section line.

7.4.2 Gradient Methods

A vivid interpretation of an optimization problem may be obtained by imagining a man standing at a slope of a valley covered by thick forest. He wishes to go to the bottom of the valley, but cannot see the entire valley—what he can survey is a limited area around him. Under the circumstances it is most natural to move in the direction of steepest slope, i.e. in the direction opposite to that of the gradient of q (x). The resultant strategy, called a *gradient method*, is a sequence of steps each of which involves two operations:

(1) Determine the direction of steepest slope, i.e. the direction of the antigradient of q (x).

(2) Move in this direction a specified distance.

The character of a trajectory in gradient search is shown in Fig. 7.6*b*.

Mathematically, a strategy of the gradient method may be obtained by making the distance $\Delta x^{(i)}$ moved at step i along each axis proportional to the component $\partial q/\partial x^{(i)}$ along this axis,

$$\Delta x^{(i)} = -\gamma \frac{\partial q\,(\mathbf{x})}{\partial x^{(i)}}, \quad i = 1, \ldots, n \tag{7.74}$$

This means that matrix Γ_n in (7.73) will be diagonal with elements γ in the principal diagonal,

$$\Gamma_n = \gamma \mathbf{I} \tag{7.75}$$

The correction at the nth step may be given in the form

$$\Delta \mathbf{x}_n = -\gamma \mathbf{I} \nabla q\,(\mathbf{x}_n) \quad -\gamma \nabla q\,(\mathbf{x}_n) \tag{7.76}$$

The strategy defined by this expression describes motion with steps of variable length, as the size of steps depends on the value of ∇q (x). Since the gradient decreases on flat slopes and near the optimum point, steps will be small at some sections thus increasing the time of search.

This setback can be eliminated by invoking a strategy with steps of permanent width. Then the correction at each step

$$\nabla x_n = -\gamma \mathbf{D}(\mathbf{x}_n) \tag{7.77}$$

may be obtained from (7.76) by replacing the gradient vector ∇q (x) with the vector of gradient direction

$$\mathbf{D}(\mathbf{x}) = \nabla q(\mathbf{x}) / |\nabla q(\mathbf{x})| \tag{7.78}$$

where

$$|\nabla q(\mathbf{x})| = \left[\left(\frac{\partial q}{\partial x^{(1)}}\right)^2 + \ldots + \left(\frac{\partial q}{\partial x^{(n)}}\right)^2\right]^{1/2} \tag{7.79}$$

Unfortunately, this gradient method is not free from disadvantages. In order to move all the time in the direction of the gradient the steps should not be very large, hence too many steps may be required. Because each step involves a new evaluation of gradient direction, which takes time, the approach to the optimum point can be very slow.

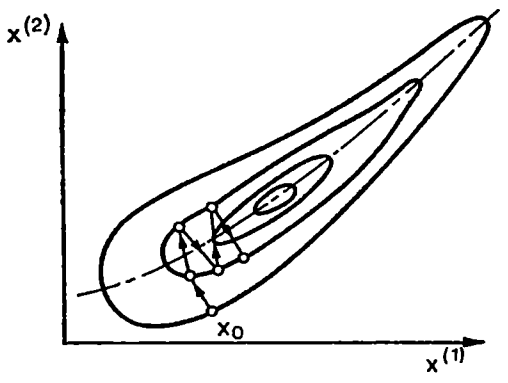


Fig. 7.7. Gradient search along the valley

Additional pitfalls occur if the optimum point $\mathbf{x}_0$ lies in a valley or on a long narrow ridge, as shown in Fig. 7.7. In such situations the gradient method may cause jumps from one slope of the valley to another and back, so that the optimum will be found after many inefficient steps. There are gradient techniques, however, that are free from these disadvantages.

7.4.3 The Method of Steepest Descent

Unlike the above gradient method in which the gradient is determined at each step, the method of steepest descent evaluates the gradient at the initial point only and moves in the found direction in equal steps so long as the magnitude of q (x) is decreasing. If q (x) is found to increase at some step, then motion in this direction terminates, the last step is eliminated altogether and a new gradient of q (x) is computed, i.e. a new direction of movement is deter-

mined. Figure 7.8*a* shows the character of trajectory in steepest descent, and the diagram at (*b*) is the algorithm of this method.

The chief advantage of the method of steepest descent is its simplicity, as most steps involve only evaluation of the magnitude of

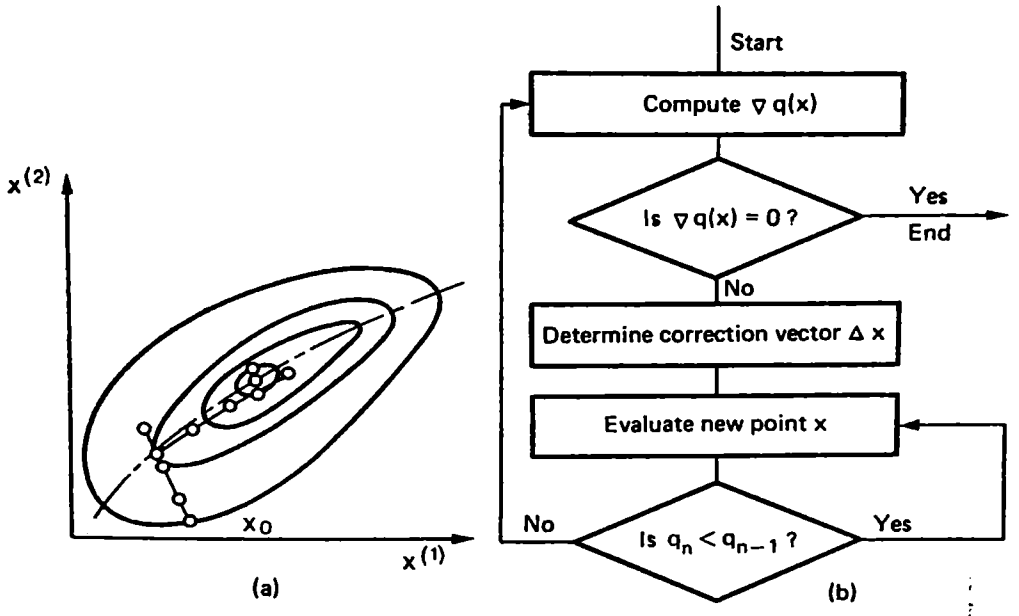


Fig. 7.8. (*a*) A trajectory and (*b*) algorithm in the method of steepest descent

q (x) rather than computation of all elements of the gradient vector However, the step has to be short so as not to go over the optimum point in the given direction, consequently the total number o steps will be significant.

7.4.4 Newton's Algorithm

On occasions where the response surface is described by a second order equation, the number of steps may be cut off appreciably with the aid of Newton's algorithm. The representation of q (x) in the form (7.66) will be sufficiently accurate even for points rather distant from the optimum point, so that matrix A itself may be taken in place of Γ_n. However, the elements a_{kj} of A computed at the optimum point are not known in advance. Nonetheless with a sufficiently good response surface, the second derivatives of q (x)

$$\frac{\partial^2 q\ (\mathbf{x})}{\partial x^{(k)}\ \partial x^{(j)}}; \quad k,\ j = 1,\ \ldots,\ n \tag{7.80}$$

taken at an arbitrary point $\mathbf{x} = \mathbf{x}_n$ will be close to the elements a_{kj} of matrix A. Defining the matrix of the second derivatives of q (x) at $\mathbf{x}_n$ by $\nabla^2 q\ (\mathbf{x}_n)$ and using it as matrix Γ_n, we obtain the correction

vector for Newton's algorithm

$$\Delta \mathbf{x}_n = -[\nabla^2 q(\mathbf{x}_n)]^{-1} \nabla q(x_n) \tag{7.81}$$

If the expansion (7.66) is an exact representation of $q(\mathbf{x})$, then the point of optimal solution is achievable in one step.

7.4.5 Multiple Extrema

The constraints of NLP problem give rise to hypersurfaces that bound the feasible region. This produces difficulties in application of methods where motion to an optimal point occurs in the direction of the gradient, as this direction may lead outside of the feasible space, as shown in Fig. 7.9. However, to reach the optimal point one should not necessarily move in the direction of the gradient. Movement may be in any direction so long as $q(\mathbf{x})$ is decreasing. Therefore, if motion meets the boundary of the feasible region, then further motion

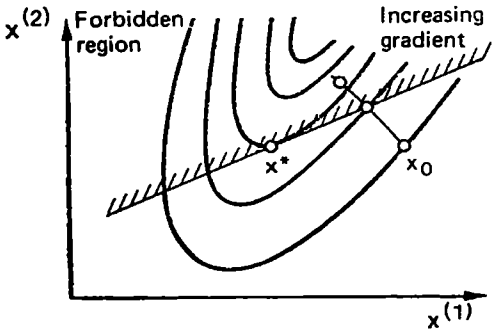


Fig. 7.9. Motion along the boundary of the feasible region

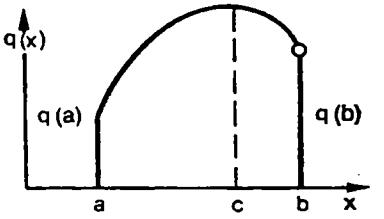


Fig. 7.10. False optimum on the boundary of the feasible region

should be along this boundary accompanied by a search for an optimal point by one of the methods, as shown in Fig. 7.9. Once an optimal point has been determined on the boundary of the feasible region, the direction of gradient should again be determined. It may happen that this direction will indicate a path inside the feasible region.

Motion along the boundary of the feasible region may be carried out also without a search algorithm, if $q(\mathbf{x})$ has been modified so that its values sharply increase outside the feasible region. Then the boundary of the feasible region will have the form of a valley bottom and if the algorithm used allows motion along the valley, then it will be along the boundary in the desired direction.

Another pitfall of constrained problems is due to motion along the boundary which may lead to a local minimum only. An example of such a case for a one-dimensional problem subject to the constraint

$a \leqslant x \leqslant b$ is given in Fig. 7.10. If the initial point is taken to the right of point c, then any of the considered methods will lead to $x = b$ where $q(x)$ attains only a relative minimum. The absolute minimum is at $x = a$ and it can be attained only by starting the motion from points lying on the left from c. This type of problem is referred to as problems with multiple extrema.

Multiple extrema problems are not necessarily associated with the presence of constraints. False optima may be due to a complex form of $q(\mathbf{x})$, as shown in Fig. 7.11.

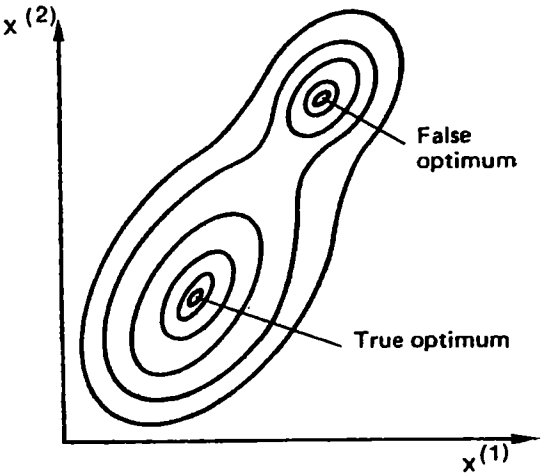


Fig. 7.11. Biextremal objective function

Because all considered techniques refer to the class of local methods, when the direction and stepwidth of the next step is evaluated by investigating a limited area of $q(\mathbf{x})$ near the point reached at the preceding step, the possibility to attain a global optimal point is entirely dependent on how successful the choice of the initial estimate has been. In these conditions it is desirable to organize a multiple search of extremum starting from different points. One of algorithms realizing this idea is described below.

The initial point $\mathbf{x} = \mathbf{x}_1$ is selected at random and from this point we seek for a local optimum $q_1 = q(\mathbf{x}_1^*)$. Keeping these values of x_1^* and q_1 in store, we restart the process by selecting at random another initial point $\mathbf{x} = \mathbf{x}_2$ and find the respective local optimum $q_2 = q(\mathbf{x}_2^*)$. If $q_2 < q_1$, then q_1 and $\mathbf{x}_1^*$ give their place in the store to q_2 and $\mathbf{x}_2^*$. If $q_2 > q_1$, then we deem the choice of $\mathbf{x}_2$ unsuccessful and restart the process. In this way we keep in store the better of found values of $q(\mathbf{x}^*)$.

The above procedure is monotonous and the memorized values of the function cannot increase. Because the initial point is selected at random at each stage, we may expect that after a rather large number of stages we succeed in visiting all local extrema and in selecting best of them.

However, in practical problems the time allocation for search is always limited, therefore a result of optimization is deemed satisfactory if a few later attempts failed to diminish $q(\mathbf{x})$.

A more elaborate discussion of various techniques of search of extrema may be found in books by Traub [48] and by Pervozvansky [49].

EXERCISES

7.1. Transform the inequalities (7.7) into equation. What is the physical sense of the additional variables?

7.2. Determine by the simplex method an optimal allocation of resources by the data of Example 7.5 taking x_1, x_2, and x_3 as nonbasic variables at the initial step.

7.3. Use the simplex method to solve the dual problem of Example 7.5 assuming u_1, u_2, u_5, and u_7 to be nonbasic variables. Compare the matrix of the optimal solution to that of Exercise 7.2.

8

DYNAMIC PROGRAMMING

8.1 Optimal Control as Variational Problem

8.1.1 Formulation of Optimal Control Problem

A typical trend of control system development is to obtain systems that are best in a certain sense. In process control and system management this trend manifests itself in manufacturing maximum number of items under limited resources (raw materials, energy, etc.). In automatic motion control systems, the principle is to minimize the time of arriving at a given point of space, or at a given trajectory under constraints on deviation of rudders, amount of burnt fuel, etc. In servo systems and stabilizing systems, the objective is to achieve maximum accuracy in circumstances with a variety of constraints imposed on the coordinates of the controlled system, actuators and the regulator.

In all these examples, the control and management problem reduces to the evaluation of a process that is best in a certain sense. Such problems relate to the class of dynamic control problems.

In Chapter 6 we have learned that the mathematical formulation of dynamic prolems of optimal control is as follows. Given a controlled system, or process whose state is described by a multivariate $\mathbf{x} = (x^{(1)}, \ldots, x^{(N)})$. The character of processes in the system can be altered by using a control u from the space of admissible controls, U, called the *control region*. In the general case, a control $u \in U$ may be also a multidimensional quantity, $\mathbf{u} = (u^{(1)}, \ldots, u^{(R)})$.

The behavior, or the motion of the controlled system is described by the system of differential equations

$$\begin{aligned} \dot{x} &= g(x, u) \\ x(0) &= c \end{aligned} \tag{8.1}$$

The measure of effectiveness of control is usually taken in the integral form

$$J_1(u) = \int_0^T Q_1[x(t), u(t)]\, dt \tag{8.2}$$

This function has the physical sense of losses, with T being the interval of the control process, and $Q_1[x(t), u(t)] = q_1(t)$ the instan-

taneous losses at time i under control $u(t)$ applied to the system in state $x(t)$.

The problem may have also constraints imposed on resources or domains of certain parameters, expressed by the relation

$$\int_0^T H[x(t),\ u(t)]\,dt = K \tag{8.3}$$

that follows from (6.50).

In Chapter 6 we established that a control $u \in U$ for which the performance index (8.2) assumes the smallest (largest) value subject to the constraints (8.3) is called an *optimal control.*

The problem of optimal control in the above statement relates to the class of variational problems handled by the calculus of variations. The quantity $J_1(u)$ defined by (8.2) is called a functional (see Sect. 2.5). Unlike a function, $f(x)$, whose numerical values are given on the set of argument x, the domain of $f(x)$, the numerical values of the functional $J_1(u)$ are given on the set of all admissible controls. The problem of optimal control is to find from the set of admissible controls, i.e. control region U, a control that would minimize the functional $J_1(u)$.

As a rule the problems of minimization of functional (8.2) slaved to a differential relationship (8.1) and subject to the integral constraint (8.3) are replaced by the minimization of another functional

$$J(u) = \int_0^T Q_1(x,\ u)\,dt + \lambda \int_0^T H(x,\ u)\,dt \tag{8.4}$$

that is slaved only to the differential relation (8.1). The parameter λ, called a Lagrange multiplier, plays the role of limited resources in process management problems. Its value is found from the boundary conditions of the variational problem.

A possibility of simplifying the variational problem with integral constraints by introducing Lagrangian multipliers follows from the following theorem.

Theorem 8.1. If $u(t)$ is an optimal control for which the functional (8.4) reaches an absolute minimum subject to the constraints (8.3), then for this $u(t)$ the functional (8.2) also achieves an absolute minimum subject to (8.3).

Proof. We construct the proof by contradiction. Let $v(t)$ be another control different from $u(t)$ and such that

$$\int_0^T Q_1(x,\ v)\,dt < \int_0^T Q_1(x,\ u)\,dt \tag{8.5}$$

and the condition

$$\int_0^T H(x, v)\,dt = K \tag{8.6}$$

is satisfied. Then,

$$\int_0^T Q_1(x, v)\,dt + \lambda \int_0^T H(x, v)\,dt = \int_0^T Q_1(x, v)\,dt + \lambda K$$

$$< \int_0^T Q_1(x, u)\,dt + \lambda K = \int_0^T Q_1(x, u)\,dt + \lambda \int_0^T H(x, u)\,dt \tag{8.7}$$

which contradicts the initial assumption of $u(t)$ minimizing (8.4).

Unfortunately the methods of variational calculus have not become a widely accepted mathematical tool for handling optimal control problems owing to some difficulties associated with these methods. Therefore we shall not dwell on the solution techniques for the variational problems, referring the interested reader to text on the calculus of variations, specifically, to the text of Elsgolts [50]. Some points important for the subsequent discussion will be noted below.

An important concept of the calculus of variations is the one of the variation of a function, which for a functional plays the same role as the differential in investigation of functions.

Let $f(x)$ be a function continuous on the interval $[a, b]$. Consider an internal point of this interval and some fixed increment in the argument, Δx. The difference or increment in f

$$f(x + \Delta x) - f(x) = df(x) = f'(x)\Delta x \tag{8.8}$$

is called the differential of f at x. As will be recalled the condition $df(x) = 0$ is necessary for $f(x)$ to be a minimum at x.

To obtain an analogy with the calculus of variations it is convenient to define $df(x)$ in a somewhat different way. Consider $f(x + \varepsilon\Delta x)$ which is a function of ε for fixed x and Δx. Differentiating this function with respect to ε and letting $\varepsilon = 0$, we obtain

$$\frac{\partial}{\partial \varepsilon} f(x + \varepsilon\Delta x)\,|_{\varepsilon=0} = f'(x)\,\Delta x = df(x) \tag{8.9}$$

Let us look at the similar concepts of the calculus of variations. Let $u(t)$ and $u_1(t)$ be two controls. The difference

$$\delta u(t) = u_1(t) - u(t) \tag{8.10}$$

is called the *variation of* $u(t)$ and the difference

$$\delta J(u) = J(u + \delta u) - J(u) \tag{8.11}$$

the *variation of the functional.*

The variation of functional may be defined in another way. Consider the functional

$$J(u + \varepsilon\delta u) = \varphi(\varepsilon) \tag{8.12}$$

that is a function of ε at fixed $u(t)$ and $\delta u(t)$. Assume that the functional is defined for different ε. This implies that different controls $u(t) + \varepsilon\delta u(t)$ are possible near the fixed value of control $u(t)$, i.e. $u(t)$ is an internal point of the control region U. The variation of functional may be defined using a similar argument to that for the differential of a function (8.9), namely,

$$\delta J(u) = \frac{\partial}{\partial \varepsilon} J(u + \varepsilon\delta u)\,|_{\varepsilon=0} \tag{8.13}$$

If $u(t)$ is an optimal control, then $\varphi(\varepsilon)$ will be a minimum at $\varepsilon = 0$. Then,

$$\frac{\partial \varphi(\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} J(u + \varepsilon\delta u)\Big|_{\varepsilon=0} = 0 \tag{8.14}$$

that is $\delta J(u) = 0$. Thus, an optimal control is the one for which the variation of the functional vanishes.

In the calculus of variations the condition $\delta J(u) = 0$ is used to obtain Euler's equation. In the set of solutions of this equation a control $u(t)$ is evaluated that minimizes the functional (8.4).

8.1.2 Solution of Variational Problems

A number of pitfalls involved in the solution of the variational problem of optimal control are of principal character.

a. The variational methods are capable of evaluating relative extrema of $J(u)$ only, whereas it is the absolute maximum or minimum that is desired.

b. For many problems of engineering value Euler's equations are nonlinear, thus preventing an explicit solution of the variational problem.

c. Constraints imposed on the magnitude of control signals prevent the search of optimal solution from being carried out by variational methods.

Because the last circumstance was of decisive importance for the development of new ideas in the field of optimal control, we discuss it in little more detail.

Common constraints imposed on control signals are limitations in the form

$$|u_i(t)| \leqslant M_i \tag{8.15}$$

implying that the magnitude of control signals fed to the actuators must be limited by magnitude (for example, the maximum voltage fed across the armature of an electric motor, or the ultimate angle of

the rudder in an airplane, or the maximum temperature in the combustion chamber of a jet plane, etc.). On the other hand, the optimal behavior of a process of plant requires as a rule that the control signals be kept at their extreme values, as this pattern of control corresponds to the fastest and efficient mode for the processes in the controlled system. A typical profile of optimal control $u\ (t)$ is illustrated in Fig. 8.1.

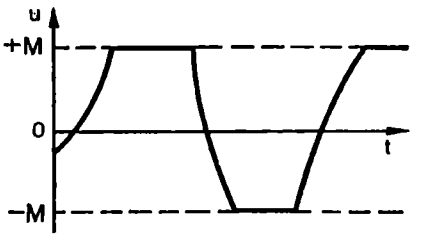


Fig. 8.1. Typical form of an optimal control signal

It is important to emphasize, however, that the extreme values of control $u\ (t)$ lie on the boundaries of the control region U, that is, are not internal for the region where variational methods are applicable. The limitation (8.15) may be avoided by introducing new variables v_i related to u_i by $u_i = M_i \sin v_i$. Then the values $|u_i| = M_i$ will correspond to $v_i = \pm \pi/2$ which are internal points of the new control region. However, this change of variables usually leads to much more complicated equations.

Much research effort has been given to solution of optimal control problems in the Soviet Union and in the United States. Soviet mathematician L.S. Pontryagin and his scholars V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko developed a theory of optimal control based on Pontryagin's maximum principle. This principle put optimal control theory on a rigorous mathematical footing and opened up wide possibilities for its practical application in the field of automatic control systems. We refer the interested reader to discussions of general control theory in Refs. [5, 36, 51-53].

Another approach to optimal process computation, called dynamic programming [54-56], will be discussed in detail below. This method provides the analyst with an efficient computational procedure for control and management optimization, which lends itself well to computer implementation.

8.2 Multi-Stage Decision Problems

8.2.1 Discrete Form of the Variational Problem

The pitfalls intrinsic to the variational problem are being overcome by means of efficient computational techniques of which one is dynamic programming. This method has been devised for the search of optimal control in multi-stage problems. However, it is used to tackle variational problems as well if these are represented in discrete form.

With reference to Theorem 8.1 we formulate the variational problem as follows. For a controlled system described by the differential equation (8.1), find a control $u(t)$ from the control region U, such that minimizes the functional

$$J(u) = \int_0^T Q[x(t),\ u(t)]\,dt \tag{8.16}$$

where

$$Q(x,\ u) = Q_1(x,\ u) + \lambda H(x,\ u) \tag{8.17}$$

If we assume that $u(t)$ is applied at discrete time instants only, i.e. that $t = k\delta$, $k = 0, 1, \ldots, n-1$; $\delta = T/n$, then we obtain a discrete formulation of our problem. Now, the functions $x(t)$ and $u(t)$ will be replaced with the sequences of

$$x_k = x(t)\,|_{t=k\delta} \text{ and } u_k = u(t)\,|_{t=k\delta}$$

Substituting the ratio $(x_{k+1} - x_k)/\delta$ for the derivative $\dot{x} = dx/dt$ in (8.1) transforms Eq. (8.1) into the difference equations

$$(x_{k+1} - x_k)/\delta = g(x_k,\ u_k) \tag{8.18}$$

Usually this equation is written in a more convenient form

$$\begin{gathered} x_{k+1} = x_k + g(x_k,\ u_k)\delta \\ k = 0,\ 1,\ \ldots,\ n-1;\ x_0 = c \end{gathered} \tag{8.19}$$

The integral (8.16) gives way to the sum

$$J_n(u) = \sum_{k=0}^{n-1} Q(x_k,\ u_k)\,\delta \tag{8.20}$$

where u represents the sequence of controls used

$$u = (u_0,\ \ldots,\ u_{n-1}) \tag{8.21}$$

In this formulation the problem seeks for controls $u_0, u_1, \ldots, u_{n-1}$ such that would minimize the sum (8.20).

In many control problems it is convenient to let $\delta = 1$. Specifically, this is the case when the process naturally partitions into separate stages or steps so that within each stage $u(t)$ remains unchanged. This leads us to a multi-stage control process considered in Chapter 6, where x_k and u_k represent the state of the controlled system and the control applied at the outset of each stage. Define

$$x_k + g(x_k,\ u_k) = T(x_k,\ u_k) \tag{8.22}$$

then we may write (8.19) in the form

$$x_{k+1} = T(x_k,\ u_k),\ k = 0,\ 1,\ \ldots,\ n-1;\ x_0 = c \tag{8.23}$$

that coincides with (6.65) which governs the transformation of the state in one step of the multi-stage control process. For $\delta = 1$, the sum (8.20) becomes

$$J_n(u) = \sum_{k=0}^{n-1} Q(x_k, u_k) \tag{8.24}$$

that coincides with (6.67) for the performance measure in a multiple-stage control process.

In what follows we shall discuss the method of dynamic programming as applied to the multi-stage control or decision process. The relationship established between the formulations of the discrete variational problem and the multi-stage control process allows us to expand the above results to the solution of variational problems.

8.2.2 Recurrence Relation

Optimization of an n-stage process consists in evaluating a sequence of controls $u_0, u_1, \ldots, u_{n-1}$ such that minimizes the performance measure $J_n(u)$. This minimal value of the performance measure will depend only upon the initial state x_0 and may be denoted

$$f_n(x_0) = \min_{u_0} \min_{u_1} \ldots \min_{u_{n-1}} [Q(x_0, u_0) + Q(x_1, u_1) + \ldots + Q(x_{n-1}, u_{n-1})] \tag{8.25}$$

Note that the first term of this expression depends only upon the control u_0, while the other terms depend on both u_0 and the control applied on other steps. For example, $Q(x_1, u_1)$ depends on u_1 and on u_0 since $x_1 = T(x_0, u_0)$. A similar argument holds for other addends. Therefore (8.25) may be written as

$$f_n(x_0) = \min_{u_0} \{Q(x_0, u_0) + \min_{u_1} \ldots \min_{u_{n-1}} [Q(x_1, u_1) + \ldots + Q(x_{n-1}, u_{n-1})]\} \tag{8.26}$$

If we note that the expression

$$\min_{u_1} \ldots \min_{u_{n-1}} [Q(x_1, u_1) + \ldots + Q(x_{n-1}, u_{n-1})] \tag{8.27}$$

represents the minimum value of the performance measure for the $(n-1)$-stage process with the initial state x_1, then in agreement with (8.25) we may denote this quantity by $f_{n-1}(x_1)$, so that

$$f_n(x_0) = \min_{u_0} [Q(x_0, u_0) + f_{n-1}(x_1)] \tag{8.28}$$

Repeating this argument for the initial state x_1 and the $(n-1)$-stage process, we obtain the minimum value of the performance

measure in the form

$$f_{n-1}(x_1) = \min_{u_1} [Q(x_1, u_1) + f_{n-2}(x_2)] \tag{8.29}$$

Continuing this argument, we write a similar expression for an $(n - l)$-stage process that starts at x_l, viz.,

$$f_{n-l}(x_l) = \min_{u_l} [Q(x_l, u_l) + f_{n-(l+1)}(x_{l+1})] \tag{8.30}$$

This equation, often referred to as Bellman's equation, is a recurrence relation yielding in succession an optimal control at each stage of the process.

The idea to optimize the control process at each step, when the process is hard to optimize as a whole, is not original and is widely accepted in practice. It is often overlooked, however, that optimization at each stage does not imply the optimization of the process as a whole. Offering up a piece is never advantageous from the standpoint of a move in a chess game, but may be advantageous from the standpoint of the whole game. Expenditures for equipment do not bring an immediate gain, but they are advantageous for the activity in the long run.

A specific feature of dynamic programming is that the method combines the simplicity of optimization at a separate step with the perspective that takes into account the far consequences of the particular step.

The choice of a control or decision at a separate step is governed not by the interests of this step, expressed as minimization of losses at this step, i.e. $Q(x_l, u_l)$, but by the interests of the entire process, expressed as the minimization of the total losses $Q(x_l, u_l) + f_{n-(l+1)}(x_{l+1})$ at all the subsequent steps. This leads us to the basic property of an optimal process, called the *principle of optimality*, which states that *whatever the initial condition and control may be, the remaining controls or decisions must constitute an optimal policy with regard to the state resulting from the first control action or decision.*

The immediate implication of this principle is that the optimization of control for an arbitrary stage of a multi-stage process consists in choosing the subsequent controls only. Therefore it is convenient to deal with the future decisions, rather than the past, to lead the process into the final state. From this standpoint, Eq. (8.30) should be rewritten in another form.

Define the number of steps $n - l$ to the end of the process to be k, then the quantities $x_l = x_{n-k}$ and $u_l = u_{n-k}$ will be written simply as x and u and imply the state of the system and the control at k steps to the end of the process. Denote also by x' the state to which the system is carried from state x by control u (in the former notation this

is $x_{n-(l+1)}$). Now, Eq. (8.23) will be written as

$$x' = T(x, u) \tag{8.31}$$

and the recurrence relation (8.30) becomes

$$f_k(x) = \min_u [Q(x, u) + f_{k-1}(x')] \tag{8.32}$$

8.2.3 Computational Aspects

Dynamic programming is a numerical method of optimization-problem solving and therefore involves fairly unwieldy computations. This, however, is not an obstacle with the capacity of modern computers.

At an arbitrary step k, counted from the final state backwards, the optimal control is defined by (8.31) and (8.32) which will be rewritten for convenience of computations. Denote by $F_k(x, u)$ the performance criterion of a k-stage process guided by optimal controls at the last $k - 1$ steps and initiated by an arbitrary control action at the initial step. Then (8.32) may be written as

$$F_k(x, u) = Q(x, u) + f_{k-1}(x') \tag{8.33}$$

$$f_k(x) = \min_u F_k(x, u) \tag{8.34}$$

where x' is determined from (8.31).

The variables x and u may take on values from finite sets or may vary continuously in some range. In the last case we should quantize these variables by sampling finite sets of possibly equidistant values. Consequently in both cases, x and u may be viewed as elements of finite sets $X = \{x^{(1)}, \ldots, x^{(m)}\}$ and $U = \{u^{(1)}, \ldots, u^{(r)}\}$. The value $Q(x, u)$ is convenient to compute in advance and specify as a table on the direct product of sets $X \times U$. This table is stored in the computer.

Table 8.1 is filled with the aid of (8.33) and (8.34). The first two columns list all possible combinations of $x \in X$ and $u \in U$. The last columns have the minimum value of $F_k(x, u)$ listed for each $x \in X$, i.e. determine $f_k(x)$ and optimal controls u^*.

Similar computations are carried out for each step of the multi-step decision process. To determine $f_k(x)$ one should first secure the table for $f_{k-1}(x)$ as this is needed to find the values of $f_{k-1}(x')$ in filling Table 8.1. Consequently, the values of $f_{k-1}(x)$ should be computed in advance of $f_k(x)$. It follows then that the computation of $f_k(x)$ must begin with the last step of the multi-step process. If we note that $f_0(x) = 0$, then the initial step should be $k = 1$ for which

$$F_1(x, u) = Q(x, u) \text{ and } f_1(x) = \min Q(x, u) \tag{8.35}$$

Table 8.1

x	u	$x' = T(x, u)$	$Q(x, u)$	$f_{k-1}(x')$	$F_k(x, u)$	$f_k(x)$	u^*
	$u^{(1)}$	$T(x^{(1)}, u^{(1)})$	$Q(x^{(1)}, u^{(1)})$	$f_{k-1}[T(x^{(1)}, u^{(1)})]$	$F_k(x^{(1)}, u^{(1)})$	$f_k(x^{(1)})$	u_1^*
$x^{(1)}$	...	...	...	...	...	...	
	$u^{(r)}$	$T(x^{(1)}, u^{(r)})$	$Q(x^{(1)}, u^{(r)})$	$f_{k-1}[T(x^{(1)}, u^{(r)})]$	$F_k(x^{(1)}, u^{(r)})$		
	$u^{(1)}$	$T(x^{(2)}, u^{(1)})$	$Q(x^{(2)}, u^{(1)})$	$f_{k-1}[T(x^{(2)}, u^{(1)})]$	$F_k(x^{(2)}, u^{(1)})$	$f_k(x^{(2)})$	u_2^*
$x^{(2)}$	...	...	...	...	...	...	
	$u^{(r)}$	$T(x^{(2)}, u^{(r)})$	$Q(x^{(2)}, u^{(r)})$	$f_{k-1}[T(x^{(2)}, u^{(r)})]$	$F_k(x^{(2)}, u^{(r)})$		
$x^{(3)}$	...	...	...	...	...	...	...

Further on the calculation proceeds in a usual manner for $k = 2, 3, \ldots, n$.

Once the computations have been completed and tables for $f_k(x)$ and u^*, $k = 1, \ldots, n$, compiled, we start the search for an optimal control for the entire process under the given initial conditions x_0. Using the table for $f_n(x)$, we find u_0^* corresponding to the given x_0 and compute $x_1 = T(x_0, u_0^*)$. Then with the table for $f_{n-1}(x)$ we find u_1^* corresponding to the found x_1 and calculate $x_2 = T(x_1, u_1^*)$, and so on. The resultant sequence is an optimal control $u^* = (u_0^*, u_1^*, \ldots, u_{n-1}^*)$.

If u is a continuous quantity, then $F_k(x^{(i)}, u)$ will be a continuous function of u. However, Table 8.1 lists but the discrete values $F_k(x^{(i)}, u^{(j)})$, $j = 1, \ldots, r$, of this function. The minimal value $F_k(x^{(i)}, u)$ may occur at an intermediate value of u. In this case u^* and $f_k(x^{(i)})$ should be determined by interpolation formulas. The interpolation formulas are needed also for the evaluation of an optimal control u_k^* when the values $x_k = T(x_{k-1}, u_{k-1})$ lie between the x values listed in Table 8.1.

Example 8.1. *A climbing problem.* An airplane flies at a speed of v_0 at an altitude h_0. It is desired to change its velocity to v_1 and altitude to h_1 so that the amount of fuel spent for this maneuver is a minimum.

The process of changing v and h may be represented on the v, h plane. We quantize the variables by breaking the variation intervals for v and h into four subintervals. The discrete states of the plant will be represented by the nodes of the mesh shown in Fig. 8.2. We assume further that only two controls are admissible at each node, $u_i = 0$ representing the variation of speed only, and $u_i = 1$, variation of altitude only. The control region, therefore, will be $U = \{0, 1\}$.

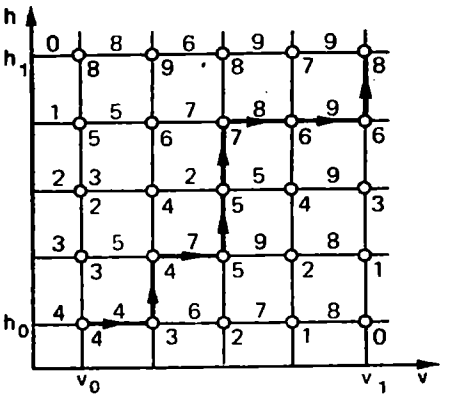


Fig. 8.2. Trajectory in the climb problem

Figure 8.2 shows one of possible trajectories corresponding to the control $u = (01011001)$. In order to estimate the quality of this trajectory we need to know the amount of fuel burnt at each step. This will be the objective function $Q(x, u)$ whose values will be given in the form of conditional figures, indicated next to each transition in Fig. 8.2. The total amount of fuel for the trajectory traced in Fig. 8.2. is summed up from the values of the performance criterion as $4 + 4 + 7 + 5 + 7 + 8 + 9 + 8 = 52$.

In order to represent this process as a multi-stage one we assign to the discrete values of v numbers from 0 to 4 starting with the

terminal value. Similar marking will be given to h. Then x_{ij} will represent a state at $v = i$ and $h = j$ with $i + j$ steps to go to the end of the process.

Denote by X_k the set of states from which the process is k steps to go, i.e. constituted by all x_{ij} for which $i + j = k$. Letting $k = 0, 1, 2$, etc., we get $X_0 = x_{00}$, $X_1 = \{x_{10}, x_{01}\}$, $X_2 = \{x_{20}, x_{11}, x_{02}\}$, etc.

Now we are in a position to solve the problem. For $k = 1$ we have $X_1 = \{x_{10}, x_{01}\}$, $F_1(x, u) = Q(x, u)$, and $f_1(x) = \min_u F_1(x, u)$.

These relations will be used to compile Table 8.2 for the calculation of the optimal control at the last step, arrayed in a pattern such as Table 8.1.

Table 8.2

x	u	x'	$Q(x, u)$	$f_1(x)$	u_1^*
x_{10}	0	x_{00}	9	9	0
	1	—	—		
x_{01}	0	—	—	8	1
	1	x_{00}	8		

For $k = 2$ we have $X_2 = \{x_{20}, x_{11}, x_{02}\}$, $F_2(x, u) = Q(x, u) + f_1(x')$, and $f_2(x) = \min_u F_2(x, u)$.

These relations will be used to compile Table 8.3, the calculation of the optimal control at the penultimate step. Similar computations

Table 8.3

x	u	x'	$Q(x, u)$	$f_1(x')$	$F_2(x, u)$	$f_2(x)$	u_2^*
x_{20}	0	x_{10}	9	9	18	18	0
	1	—	—	—	—		
x_{11}	0	x_{01}	9	8	17	16	1
	1	x_{10}	7	9	16		
x_{02}	0	—	—	—	—	14	1
	1	x_{01}	6	8	14		

are carried out at $k = 3$, 4, etc. (see Table 8.4, the computation of the optimal control with three steps to go).

Table 8.4

x	u	x'	$Q(x, u)$	$f_3(x')$	$F_3(x, u)$	$f_3(x)$	u_3^*
x_{30}	0	x_{20}	6	18	24	24	0
	1	—	—	—	—		
x_{21}	0	x_{11}	8	16	24	24	0
	1	x_{20}	8	18	26		
x_{12}	0	x_{02}	9	14	23	22	1
	1	x_{11}	6	16	22		
x_{03}	0	—	—	—	—	17	1
	1	x_{02}	3	14	17		

The data of Tables 8.2-8.4 can be displayed on the plane (v, h) by showing the values $f_k(x)$ as numbers in the respective nodes, and u^* by arrows pointing toward the next node, as shown in Fig. 8.3. Once $f_k(x)$ and u^* have been determined for all nodes of the mesh, we are in a position to find an optimal trajectory, moving from the initial node in the direction shown by the arrows. The optimal control will be $u^* = (11000110)$ associated with $f_n(u^*) = 37$.

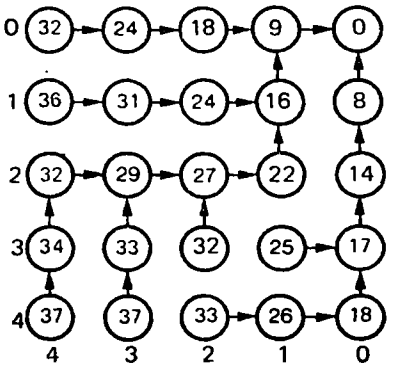


Fig. 8.3. Optimal trajectory in the climb problem

This example illustrates well the advantages of the method of dynamic programming. Firstly, this method eliminates the problem of absolute and relative minima, as from the computational procedure it should be evident that it always evaluates an absolute minimum. Secondly, the limitations of the form $|u_i| \leqslant M_i$, that have been a formidable obstacle to application of variational methods, facilitate the computational procedure of dynamic programming, as they constrain the control region U. Finally, the method of dynamic programming markedly simplifies the search of optimal solution as compared with a simple exhaustive search. This statement will be illustrated by a calculation.

If each stage admits r different controls, then in exhaustive search each of these controls should be evaluated in combination with all feasible controls at all other steps, which runs for an n-step process into r^n versions. For a modest multi-stage problem of $r = 10$ and $n = 10$ this yields 10^{10} versions.

For comparison, to choose a control at some step for state x, dynamic programming considers only those of possible continuations that correspond to optimal continuation from the state $x' = T(x, u)$. This eliminates an enormous number of unuseful versions. For example, if each stage allows m states with r controls for each state, then the total number of options to consider will be rmn or r^2n if $r = m$, i.e. if we assume that there exists a control that carries each state into any new state. For $r = 10$ and $n = 10$ this runs into only 10^3 versions.

For all its advantages, however, dynamic programming is not free from inconveniences. These are associated with the treatment of a sequence of controls or decisions being in reverse order, when the analyst has to go backward from the end of the process to its start, defining an optimal control for all feasible states at this stage, and until the very last instant remaining unaware of what the optimal control for the given initial state will be. In the final analysis he finds out that most of the calculations have been in vain, as the results of the optimal control evaluation for the states not in the optimal path have not been used. An obvious advantage in this respect is offered by the maximum principle that opens up a way for constructing one individual trajectory.

It is instructive to note that although the method of dynamic programming has been evaluated as an optimization technique for dynamic problems, i.e. processes varying in time and described by a triplet (x, u, t), it may be applied to statistical optimization problems if the role of parameter t will be relegated to some other parameter allowing partition into a sequence of stages. For example, the evaluation of the shortest path in a graph elucidated in Sect. 3.2 is but a modification of the method of dynamic programming for the statistical optimization problem. Kuzik [5] discusses other examples of this type.

8.2.4 Final State Control

In Chapter 6 we have already noted that for some problems the point that matters is the final state x_n in which the process arrives, rather than the character of motion. The criterion for such problems will be the value of the objective function at the end of the control process,

$$q = q(x_n) \tag{8.36}$$

Define as before x and u to be state of the system and the control applied with k steps remaining to go to the end of the process, and $x' = T(x, u)$ to be the next state, i.e. the one with $k - 1$ steps to go. The state x_n and hence the objective function $q(x_n)$ depend on the initial state x and on all the subsequent controls.

To arrive at a recurrence relation defining an optimal control, denote by $F_k(x, u)$ the value of the objective function $q(x_n)$ with k steps to the end of the process started at x with an arbitrary control u at this initial step and with optimal controls at the remaining $k - 1$ steps, and by $f_k(x)$ the value of $q(x_n)$ for optimal controls at k steps to the end of the process so that

$$f_k(x) = \min_u F_k(x, u) \tag{8.37}$$

The value $F_k(x, u)$ can be obtained on the lines of the following argument. An arbitrary control u carries the controlled system at the initial step from state x to state $x' = T(x, u)$. To obtain $q(x_n) = F_k(x, u)$ we need to apply optimal controls at the remaining $k - 1$ steps, which yields $q(x_n) = f_{k-1}(x') = f_{k-1}[T(x, u)]$. This will be the value $F_k(x, u)$. Thus,

$$F_k(x, u) = f_{k-1}[T(x, u)] \tag{8.38}$$

Substituting $F_k(x, u)$ in (8.37), we have

$$f_k(x) = \min_u f_{k-1}[T(x, u)] \tag{8.39}$$

This formula is a complete analog of (8.32). Applying the computational procedure outlined above to this formula and observing that

$$f_0(x) = q(x) \tag{8.40}$$

one may determine an optimal control for any initial state of a multi-stage process.

8.2.5 Recurrence Relation for Markov Processes

We consider application of the method of dynamic programming to optimization of random processes by taking as an example a controlled Markov chain.

As shown in Sect. 5.7, a Markov chain may be specified by a transition probability matrix $\mathbf{P} = [p_{ij}]$ of type $L \times L$. In order that the Markov chain be controllable there should be a possibility to alter transition probabilities p_{ij} from outside. Assume that there exists a possibility to guide a Markov process in N various ways, a way k being associated with the transition probability matrix $\mathbf{P}_k = [p_{ij}^{(k)}]$, $k = 1, \ldots, N$. We shall call each method of guiding the Markov process a *strategy*.

We assume that each method of guiding can be estimated by specifying a matrix of losses, or what is convenient in this case, a matrix of returns $\mathbf{R}_k = [r_{ij}^{(k)}]$. The quantities $r_{ij}^{(k)}$, $i, j = 1, \ldots, L$, $k = 1, \ldots, N$, give the receipt per step in moving the process from state i to j with the kth strategy.

Example 8.2. The manufacturer from Example 5.14 can increase sales by appropriate advertisement but in state 1 this would require additional expenditures and reduce the profit. If the manufacturer is in state 2, it can accelerate the transition to state 1 by increasing its research funds. Strategy 1 tends to avoid expenses for advertisement and research, while strategy 2 involves these expenditures. The transition probability matrices for these strategies may be as follows:

$$\mathbf{P}_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.6 \end{pmatrix}, \quad \mathbf{R}_1 = \begin{pmatrix} 9 & 3 \\ 3 & -7 \end{pmatrix},$$

$$\mathbf{P}_2 = \begin{pmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{pmatrix}, \quad \mathbf{R}_2 = \begin{pmatrix} 4 & 4 \\ 1 & -19 \end{pmatrix}$$

Our problem is to find a strategy providing the maximum revenue in n steps of the controlled Markov process.

Denote by $f_i(n)$ the maximum receipt in n steps started from state i. Assume that one step is made from state i to state j with revenue $r_{ij}^{(k)}$ while the other $n-1$ transitions are made in an optimal manner. The total revenue for n steps will then be $r_{ij}^{(k)} + f_j(n-1)$.

Actually, however, the first transition (i, j) occurs with probability $p_{ij}^{(k)}$. Because the first transition occurs at random we should take into account the possibility of transition to all feasible states with $j \in \{1, \ldots, N\}$. Therefore the expected revenue

$$F_i(n, k) = \sum_{j=1}^{N} p_{ij}^{(k)} [r_{ij}^{(k)} + f_j(n-1)] = q_i^{(k)} + \sum_{j=1}^{N} p_{ij}^{(k)} f_j(n-1) \quad (8.41)$$

where

$$q_i^{(k)} = \sum_{j=1}^{N} p_{ij}^{(k)} r_{ij}^{(k)} \quad (8.42)$$

is the revenue per step in moving from state i under strategy k.

The expression (8.41) includes strategy k used at the initial step. By appropriately choosing this strategy we may ensure a maximum expected revenue $f_i(n)$, which leads us to the recurrence relation

for choosing strategy k at the first step of the n-step process

$$f_i(n) = \max_k \left[q_i^{(k)} + \sum_{j=1}^{N} p_{ij}^{(k)} f_j(n-1) \right] \tag{8.43}$$

In calculating (8.43) it is convenient to array the data in tables as we have done in the preceding sections. For the problem of Example 8.2 with $n = 1$ and $n = 2$ the respective tables are Table 8.5 and Table 8.6. For $n > 2$, tables may be compiled as for the case with $n = 2$.

Table 8.5

i	k	$p_{ij}^{(k)}$		$r_{ij}^{(k)}$		$q_i^{(k)}$	$f_i(1)$	k^*
		$j=1$	$j=2$	$j=1$	$j=2$			
1	1	0.5	0.5	9	3	6	6	1
	2	0.8	0.2	4	4	4		
2	1	0.4	0.6	3	−7	−3	−3	1
	2	0.7	0.3	1	−19	−5		

Table 8.6

i	k	$q_{ij}^{(k)}$	$p_{ij}^{(k)}$		$p_{ij}^{(k)} f_i(1)$		$F_i(2, k)$	$f_i(2)$	k^*
			$j=1$	$j=2$	$j=1$	$j=2$			
1	1	6	0.5	0.5	3.0	−1.5	7.5	8.2	2
	2	4	0.8	0.2	4.8	−0.6	8.2		
2	1	−3	0.4	0.6	2.4	−1.8	−2.4	−1.7	2
	2	−5	0.7	0.3	4.2	−0.9	−1.7		

EXERCISES

8.1. Continue solving Example 8.1 and compute the entire optimal trajectory.

8.2. Treat the mesh in Fig. 8.3 as a graph to find an optimal path by using the shortest path evaluation techniques of Sect. 3.2. Compare this solution with the one obtained by the method of dynamic programming.

8.3. Given a 3-step control process of a scalar x. At the outset of the process x may be any integer within -10 to 10. The process is moved from one step to another by the transformation $x' = T(x, u) = x + u$, for which the admissible values of u are

$$\begin{cases} \{-1, 0, 1\} \text{ at step } 1 \\ \{-4, 0, 4\} \text{ at step } 2 \\ \{-9, 0, 9\} \text{ at step } 3 \end{cases}$$

The losses are given by the squared deviation of x from 0 at the end of the process, i.e. by $q(x) = (x - 0)^2 = x^2$.

Find an optimal control that leads the process in three steps to the specified end value of $x = 0$.

9

GAME THEORY

9.1 Introduction

9.1.1 Game as a Model of Conflict Situation

The theory of games is concerned with decisions, and in particular with choosing from a number of possible decisions a best one in conflict situations [42, 57, 58]. A situation will be said conflicting if it involves interests of a few, normally two, participants persueing opposite objectives. Each of the sides involved may perform actions to achieve its goals whereby a success of one side means a failure for the opponent.

The theory of games was first given a systematic development by von Neumann and Morgenstern [59] who considered mainly economic applications where the competing opponents are trading companies, industrial enterprises, and the like. However, situations involving conflict can be encountered in many other fields of life. Such are almost all military situations involving military planning, choice of weapon systems, protection of crucial installations, hunting and intercept of a target, and such. Examples of such situations are found in sport, debates, arbitration, and parlamentary elections with a few candidates for one place.

As a rule, conflict situations are hard to analyze directly as they include many factors of minor importance. In order to make a conflict situation amenable to mathematical analysis it should be simplified by keeping only the factors of principal importance. The simplified model of the conflict situation is called a *game*, and the opposing sides are called *players*. In what follows we limit our consideration to models with two opponents.

Thus, by a game we mean a set of rules which define the actions which the set of players can take, and the resulting gains or losses. This set of rules is given before a particular "play" begins. Every game can be performed in a number of possible plays. Every play consists of a series of moves made in accordance with the rules of the game. The rules tell the players the sequence of moves and the character of each move.

Each move is assigned either to a player (*personal move*) or to a chance (*random* or *chance move*). A personal move is a choice by the player of one option from the set available to the player. For example,

every move in chess is a personal move and the first one is selected from 20 available options.

A random move is also a choice of one of a set of available versions, but the choice is effected by a certain random mechanism rather than a player. Examples of random moves are drawing cards or tossing a coin. The choice of a random move is called the outcome of this move.

It is customary to lay down the rules with respect to moves as follows. For the first move, the rules specify whether this will be a personal or random move. If this is a personal move, then the rules list the available options and point out which of the players is to make a move. If this is a random move, the rules list the available versions with the probabilities of their choice.

For each subsequent move, the rules specify whether it will be a personal or a random move depending on the choices made and the outcomes of the preceding moves. If it has been a random move, then the rules list the available options and dictate the probabilities of their choice. For a personal move, the rules indicate which player is to make a move and list the available options.

Finally, the rules define, depending on the choices and the outcomes, i.e. on the course of the game, when the game is to terminate and what is the payoff for each player.

9.1.2 The Concept of Strategy

Imagine that we wish to play a game of chess with white pieces but not in person and therefore have to instruct the player who will carry out the game how he should start the game and respond to opponent's moves. The complete system of such instructions, i.e. a rule that tells the player what to do and which alternative to choose at each information set is called a *strategy*.

For example, the strategy of White should indicate the first move, then for each possible move of Black an adequate response, and so on. Evidently, compiling a complete strategy for a chess game is an enormous and an impracticable task. For example, at the very outset White is to make two decisions for the first two moves. If he plays not in person, then he has to prepare 21 options for the same two moves: one for the first move and 20 options for the 20 possible responses of Black. Nonetheless in many simpler situations the concept of strategy is rather useful.

Saying another way, a strategy of a player is a unique description of his choice at each possible situation where he is to make a personal move.

If a game consists of personal moves only, then the outcome of the game is determined if every player has chosen his strategy. When a game involves random moves, the choice of strategies by the players will not decide the final outcome of the game.

9.1.3 Two-Person Zero-Sum Games

Denote by X and Y the sets or spaces of possible strategies available to players I and II. The quantities $x \in X$ and $y \in Y$ will imply specific strategies for I and II.

In order to consider random moves as well, it is convenient to assume that there is a third player that makes random moves using the respective random mechanism. Denote by H the set of strategies of this player. Any strategy $h \in H$ representing a sequence of all random moves in a play will occur with some probability $p(h)$ which is easy to calculate if the probability of each random move in this sequence is known. It is an easy matter to see that $p(h)$ is a probability distribution on H, i.e. satisfies the conditions

$$p(h) \geqslant 0, \quad \sum_{h \in H} p(h) = 1 \tag{9.1}$$

Denote by g a version of the game, i.e. one possible play. This version will be defined if the strategies x, y of the players and the strategy of random moves h are chosen. Consequently, a specific play g is represented by the triplet

$$g = (x, y, h) \tag{9.2}$$

The result of the play is a gain or a payoff of each player. For convenience we shall assign to a payoff some monetary sum.

Consider a play $g(x, y, h)$ with $L_x(x, y, h)$ and $L_y(x, y, h)$ denoting the losses or gains of players I and II, respectively. We shall consider winnings as negative payoffs. The total sum of payoffs of both players is

$$L_x(x, y, h) + L_y(x, y, h) \tag{9.3}$$

The following discussion will be limited to *two-person zero-sum games* where payments are only made between the players, so that the sum of all their gains is zero, i.e. the loss for one is the gain for the other.

In dealing with zero-sum games one need not calculate gains or losses for both players in separate, but rather confine himself to evaluating the payoff of the second player, say,

$$L_y(x, y, h) = -L_x(x, y, h) = L(x, y, h) \tag{9.4}$$

Because the strategy h is random, for certain strategies x and y the payoff $L(x, y, h)$ will be a random variable with the probability distribution $p(h)$ on H. Accordingly, the selected strategies x and y can be estimated only by averaging the expected payoff $L(x, y, h)$ over the entire space H, i.e. by introducing the notion of expected losses $L(x, y)$ that may be determined in view of (5.52) by

$$L(x, y) = \sum_{h \in H} L(x, y, h)\, p(h) \tag{9.5}$$

A game is *finite* if all possible strategies of the players are listed, i.e. the spaces or sets X and Y are specified, and for any $x \in X$ and $y \in Y$ the payoffs are defined. We thereby arrive at a new formal definition of the game.

A game G is defined by the triplet

$$G = (X, Y, L) \tag{9.6}$$

where X and Y are some sets, and L is a bounded function defined on the direct product $X \times Y$.

Points $x \in X$ and $y \in Y$ are called the *strategies* of players I and II, and L is called the *loss* or *payoff function*.

Finite games, that is games where each player has only a finite number of possible actions to choose from, are convenient to specify in the form of a payoff matrix. Let $G = (X, Y, L)$ be a finite game with $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$. Then the $m \times n$ matrix

$$\mathbf{Q} = [q_{ij}] = \begin{pmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \vdots \\ q_{m1} & \cdots & q_{mn} \end{pmatrix} \tag{9.7}$$

In order to complete the description of the game we need to designate the goals of each player in choosing specific strategies. Player I tends to maximize his winnings, i.e. the function $L(x, y)$, while player II tends to minimize his losses, i.e. $L(x, y)$. The objectives are seen to be totally contradictory. A difficulty involved in this game is that none of the players completely controls the value of $L(x, y)$, because I controls only values of x and II controls only values of y. Overcoming this difficulty, i.e. the evaluation of a rational conduct for each player, is the essence of game theory.

It is important to emphasize that the outlined consideration holds only for zero-sum two-person games. When the number of players is more than two, then an entirely different situation occurs in which some of the players can combine their actions and form coalitions with cooperative distribution of winnings among the participants of the coalition. The members of a coalition may completely reveal their possibilities to conduct joint actions aimed at largest winnings.

Another model, particularly suitable for exonomic situations, is a *non-zero-sum game*, where the payments, i.e. gains and losses, do not add up to zero because of some payments coming from outside. These payments are considered as the losses for an additional fictitious player so as to reduce the n-person non-zero-sum game to an $n + 1$ person zero-sum game.

For $n > 2$ the theory of n-person games is rather involved and insufficiently developed, therefore we confine ourselves to the zero-sum two-person model.

Example 9.1. To illustrate the above concepts and definitions let us look at a game involving four moves.

Move 1, personal. Player I chooses one of the two integers 1 or 2.

Move 2, random. A coin is tossed and if it lands heads up (and in this case only) player II is informed about the choice of player I.

Move 3, personal. Player II chooses one of the two integers 3 or 4.

Move 4, random. One of the three integers 1, 2, or 3 is selected at random with the probabilities 0.4, 0.2 and 0.4.

The final result. The numbers chosen in the first, third and fourth moves are summed up and paid to player I if the sum is even or to player II if the sum is odd.

For the purpose of analysis the game is convenient to represent as a tree in which vertices represent the positions of the game, and

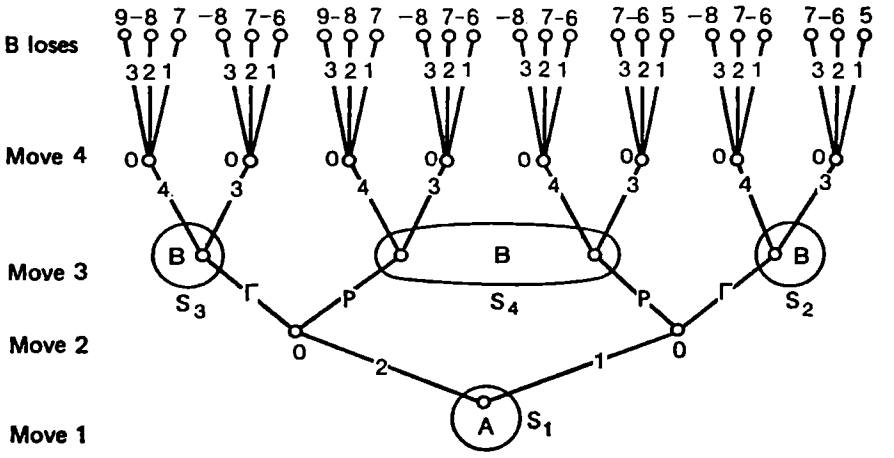


Fig. 9.1. Tree of the game

moves by arcs from one vertex to another. The tree of this game is shown in Fig. 9.1. The vertices corresponding to personal moves of players A and B are labeled by A and B, respectively. The vertices corresponding to chance moves are labeled by 0. The terminal vertices designating individual versions of the game are labeled by the payoffs of player B.

At vertices corresponding to personal moves, a player has a certain information about the preceding moves. If in some vertices the player has one and the same information, then these vertices may be combined. Such an operation leads to groups of vertices S_i called the *classes of information*. In the example under study, there are four classes of information with the contents as follows.

S_1, no move was made, player A is to make move 1;

S_2, player A has chosen 1;

S_3, player A has chosen 2;

S_4, the choice of player A is unknown.

Having arrived at a vertex of class S_4 player B has no information as to the choice of player A. Such a game is said to be performed under uncertainty. If an information class consists of only one vertex, then the player arriving at this vertex is completely certain as to all the previous moves (this information can be obtained, for example, by tracing out the path of this vertex in the tree), i.e., has a full information about the game. On the other hand, games in which every class of information contains only one vertex are said to be games with complete information (chess, draughts, dominoes).

Consider the sets of strategies of the players. The set of strategies of player A consisting of only two elements, corresponding to 1 or 2, is given below

x_1	x_2
(1)	(2)

The strategy of player B must suggest his move at any development of the game. A version of the game is determined by the class of information available to the player. Three classes of information, S_2, S_3, and S_4 are available to player B. Consequently, the strategy of player B is to suggest which of the two integers, 3 or 4, he should choose in each class of information. For example, the strategy (4, 3, 3) implies that player B chooses 4 in the information class S_2 and 3 in the classes S_3 and S_4. The set of strategies of player B is given below

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
333	334	343	344	433	434	443	444

The chance strategies h consist of two moves: the second move is to choose heads (H) or tails (T) with the probability 1/2, and the fourth move is to choose numbers 1, 2, or 3 with probabilities 0.4, 0.2, and 0.4. The probability of each strategy is equal to the product of the probabilities of the outcomes of both moves. The set of chance strategies with the associated probability distribution $p(h)$ is given below

$(H, 1)$	$(H, 2)$	$(H, 3)$	$(T, 1)$	$(T, 2)$	$(T, 3)$
0.2	0.1	0.2	0.2	0.1	0.2

In accordance with (9.5) the payoffs will be

$$q_{ij} = \sum_{h \in H} L(x_i, y_j, h)\, p(h)$$

where $L(x_i, y_j, h)$ are shown next to the terminal vertices of the tree. The calculation of q_{ij} is convenient to carry out by the table method. For the sake of illustration, Table 9.1 is constructed for the evaluation of $q_{11} = 3.6$.

Table 9.1

h	$p(h)$	$L(x_1, y_1 h)$	$L(x_1, y_1, h)\, p(h)$
$(H, 1)$	0.2	5	1.0
$(H, 2)$	0.1	−6	−0.6
$(H, 3)$	0.2	7	1.4
$(T, 1)$	0.2	5	1.0
$(T, 2)$	0.1	−6	−0.6
$(T, 3)$	0.2	7	1.4

9.1.4 Minorant and Majorant Values of the Game

In order to have an insight into the principles underlying the choice of each player of his strategy we consider a game specified by its matrix in Table 9.2.

Table 9.2

	y_1	y_2	y_3	y_4	$A(x)$
x_1	7	2	5	1	1
x_2	2	2	3	4	2
x_3	5	3	4	4	3*
x_4	3	2	1	6	1
$B(y)$	7	3*	5	6	

Assume that player I chooses strategy x_k. His winning $L(x_k, y)$ will depend on the strategy chosen by player II. For example, at strategy x_1, the payoffs to I may be 7, 2, 5, or 1. Can player I expect the maximum gain of 7? We should answer in affirmative, if II chooses strategy y_1. However, player II may choose any other strategy, including y_4 with a payoff of 1. However, no other strategy can diminish his payoff. Therefore this least element of the set $L(x_1, y) = \{7, 2, 5, 1\}$ is the guaranteed winning of I at strategy x_1. Generalizing this argument, we see that if player I chooses strategy x_k, then he is guaranteed a winning $A(x_k)$ equal to the least element of the set $L(x_k, y)$,

$$A(x_k) = \min_y L(x_k, y) \tag{9.8}$$

The theory of games assumes that the players are cautions and act without an unbased risk. Accordingly, player I should choose

a strategy $x \in X$ which corresponds to the largest of numbers $A(x)$. We denote the guaranteed gain of player I by α and call it the *minorant value* of the game or the *maximin*. In mathematical terms,

$$\alpha = \max_x A(x) = \max_x \min_y L(x, y) \tag{9.9}$$

The values $A(x)$ corresponding to the game designated by the matrix of Table 9.2 are listed in the last column, where the value of α is marked with an asterisk.

Similar argument may be carried out for player II. However, the matrix of the game lists his losses, which he tends to minimize. Consider a strategy y_k. It may result for him in a payoff not higher than

$$B(y_k) = \max_x L(x, y_k) \tag{9.10}$$

In order to secure the least payoff, player II should choose a strategy $y \in Y$ which corresponds to the smallest number of $B(y)$. We denote the payoff with which player II may limit his loss by β and call it the *majorant value* or the *minimax*,

$$\beta = \min_y B(y) = \min_y \max_x L(x, y) \tag{9.11}$$

The values $B(y)$ for the matrix game specified by Table 9.2 are listed in the bottom row of the table. The value of β is marked with an asterisk.

Theorem 9.1. If $G = (X, Y, L)$ is a game, then for any $x \in X$ and $y \in Y$,

$$A(x) \leqslant L(x, y) \leqslant B(y) \text{ and } \alpha \leqslant \beta$$

Proof. By definition

$$A(x) = \min_y L(x, y) \leqslant L(x, y) \tag{9.12}$$

$$B(y) = \max_x L(x, y) \geqslant L(x, y) \tag{9.13}$$

Hence,

$$A(x) \leqslant L(x, y) \leqslant B(y) \tag{9.14}$$

Because this relationship holds for any $x \in X$ and $y \in Y$, we may choose as x the value at which $A(x) = \alpha$ and as y the value at which $B(y) = \beta$ to get

$$\alpha \leqslant \beta \tag{9.15}$$

As can be seen, the minorant value, i.e. the winning that player I can guarantee for himself, does not exceed the majorant value, i.e. the payoff with which player II can confine himself.

9.2 Values of Games and Optimal Strategies

9.2.1 Strictly Determined Game

The simplest particular case of zero-sum games is a game with $\alpha = \beta$. Denote this value by V. This case holds in the matrix game specified by Table 9.2 where $\alpha = \beta = 3$. In this case no method of playing can guarantee player I a gain larger than β, since β is the value at which player II can limit his loss. On the other hand, no method of playing can guarantee II a payoff smaller than α, as I can guarantee a gain of α. Consequently, if $\alpha = \beta = V$, then neither strategy of any player can guarantee a result better than V. At the same time each player can have V for granted. In other words, neither I nor II have a better strategy than the one securing them with V. The value of V is called the *value of the game*, and the strategies providing this result are said to be *optimal*.

The entry in the payoff matrix that yields V is referred to as a *saddle point* since V is the largest value in the column and the smallest value in the row. The intersection of these column and row gives the saddle point. A game with a payoff table where this is so is called *strictly determined*.

A strictly determined game is called *fair* if $V = 0$, otherwise it is unfair. In order to make it fair I must pay II a value of V before each new play begins.

9.2.2 Mixed and Pure Strategies

If the game has no saddle point, its value and optimal strategies are hard to determine. Consider, for example, the matrix game designated by Table 9.3. In this game, $\alpha = 4$ and $\beta = 5$. Consequently,

Table 9.3

	y_1	y_2	$A(x)$
x_1	3	6	3
x_2	5	4	4*
$B(y)$	5*	6	

player I can guarantee a gain of 4, while player II can limit his payoff by 5. The region between β and α seems to be no one's territory and each player may try to improve his result at the expense of this region. What should be optimal strategies of these players in this case?

If either player employs the strategy marked with the asterisks (x_2 and y_1), then the winning of I and the loss of II will be 5. This leaves II at a disadvantage, as I gains more than he can guarantee for himself. However, if II somehow discovers that I intends to employ strategy x_2, then he can employ strategy y_2 that diminishes the gain of I to 4. It is true, however, that if I discovers that II intends to employ y_2, then using x_1 he may increase his gain to 6. In this situation each player should keep the strategy he intends to use a secret. However, how can one do this? If the play is repeated many times and II uses strategy y_2 all the time, then I will soon reveal the plot of II and using strategy x_1 will increase his gain. Clearly, player II should vary his strategy from one play to another, but this should be done so that I cannot discover which strategy will be used in each case.

Secrecy may be secured if a strategy is chosen each time at random with the aid of some random mechanism. For example, II may toss a coin and use strategy y_1 if it lands heads up and y_2 if it lands tails up. This tack denies the opponent of any possibility to discover in advance the strategies employed by the other side.

For the random choice mechanism, gains and losses for either player will be random variables. The result of the game may in this case be estimated by the expected payoff of player II. To illustrate, if in the matrix game of Table 9.3 player II has used strategies y_1 and y_2 chosen at random with probabilities 0.5 and 0.5, then if I chooses x_1, the expected payoff will be

$$L_{\text{exp}} = q_{11} \times 0.5 + q_{12} \times 0.5 = 3 \times 0.5 + 6 \times 0.5 = 4.5$$

and for strategy x_1 the expected payoff

$$L_{\text{exp}} = q_{21} \times 0.5 + q_{22} \times 0.5 = 5 \times 0.5 + 4 \times 0.5 = 4.5$$

Consequently, player II may limit his mean payoff by 4.5 independent of the strategy employed by player I.

In a number of situations, therefore, it is advantageous to choose a strategy at random, using a chance mechanism. A strategy based on random choice is called a *mixed strategy* in contrast to those designated in advance and called *pure strategies*. A more rigorous definition of pure and mixed strategies follows.

Let $G = (X, Y, L)$ be a game. The sets $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ listing all possible strategies are called the *sets of pure strategies*, and the elements of these sets are pure strategies.

To obtain a mixed strategy, a player should use a randomization mechanism such as coin tossing, rolling a die, and such, having the number of outcomes equal to the number of pure strategies of the player.

Denote by $\xi^{(1)}, \ldots, \xi^{(m)}$ the probabilities of outcomes of the random mechanism of player I, and by $\eta^{(1)}, \ldots, \eta^{(n)}$ the probabilities

of outcomes of the random mechanism of player II. If each outcome i of the random mechanism of I is assigned a pure strategy x_i of this player, and each outcome j of the random mechanism of II is assigned a pure strategy y_j of this player, then the ordered sets $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$ and $\eta = (\eta^{(1)}, \ldots, \eta^{(n)})$, the elements of which satisfy the normalizing conditions (5.3), will represent the probability distributions $\xi(x)$ and $\eta(y)$ on X and Y. These distributions completely define the development of the game and are referred to as *mixed strategies* of players I and II, respectively.

In the general case, each player may have a few randomization schemes with different probability distributions on X and Y, for example, $\xi_1 = (\xi_1^{(1)}, \ldots, \xi_1^{(m)})$, $\xi_2 = (\xi_2^{(1)}, \ldots, \xi_2^{(m)})$, etc.; $\eta_1 = (\eta_1^{(1)}, \ldots, \eta_1^{(n)})$, $\eta_2 = (\eta_2^{(1)}, \ldots, \eta_2^{(n)})$, etc. Then the sets

$$E = \{\xi_1, \xi_2, \ldots\} \text{ and } H = \{\eta_1, \eta_2, \ldots\} \tag{9.16}$$

will represent the sets of mixed strategies of players I and II.

9.2.3 Payoff Function in Mixed Strategies

Consider a game $G = (X, Y, L)$ in which $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ are the sets of pure strategies and $L(x, y)$ represents the payoffs of player II determined for pure strategies $x \in X$ and $y \in Y$, henceforth called the *payoff* or *loss function*.

Assume now that the players employ mixed strategies. This means that the sets $E = \{\xi_1, \xi_2, \ldots\}$ and $H = \{\eta_1, \eta_2, \ldots\}$ are specified whose elements are mixed strategies, i.e. various representations of the probabilities $\xi(x)$ and $\eta(y)$ on X and Y. The development of the play will be determined if either player has chosen his mixed strategy $\xi \in E$ and $\eta \in H$. Consequently, in using mixed strategies, the sets E and H should be substituted for X and Y in the rules of the game.

The payoff of player II will also change with mixed strategies. Because the game is randomized, the gains and losses will also be random. Therefore we may now speak of the expected winning X or payoff Y, which is defined by both the payoff function $L(x, y)$ and by the probability distributions $\xi(x)$ and $\eta(y)$ and may be evaluated with the formula for the mean value of a function of two variables

$$L(\xi, \eta) = \sum_{x, y} L(x, y)\, \xi(x)\, \eta(y) \tag{9.17}$$

The loss function $L(\xi, \eta)$ enters in the rules of the game when mixed strategies are employed.

Thus, in choosing mixed strategies we obtain a new game $\Gamma = (E, H, L)$ in place of $G = (X, Y, L)$, which may be called an averaged game G (in expectation). In this game the loss function $L(\xi, \eta)$ is determined by (9.17) as an expected value of $L(x, y)$ for the given

probability distributions $\xi(x)$ and $\eta(y)$. We summarize the above statements with the following definitions.

Let $G = (X, Y, L)$ be a game, E and H define probability distributions ξ and η over the sets X and Y, and $L(\xi, \eta)$ represent the payoff function (in expectation). Then the game

$$\Gamma = (E, H, L) \tag{9.18}$$

is called the averaging of game G.

Let $\Gamma = (E, H, L)$ be the averaging of $G = (X, Y, L)$ in expectation. Then points $x \in X$ and $y \in Y$ are called *pure strategies* in G and points $\xi \in E$ and $\eta \in H$ are *mixed strategies* in G.

It should be noted that pure strategies may be viewed as a particular case of mixed strategies. Indeed, a mixed strategy defined by the probability distribution

$$\begin{aligned} \xi(x) &= 1 \text{ for } x = x_k \\ &= 0 \text{ for } x \neq x_k \end{aligned} \tag{9.19}$$

coincides with the pure strategy x_k, and a mixed strategy defined by the probability distribution

$$\begin{aligned} \eta(y) &= 1 \text{ for } y = y_i \\ &= 0 \text{ for } y \neq y_i \end{aligned} \tag{9.20}$$

coincides with the pure strategy y_i.

9.2.4 Minimax and Maximin in Mixed Strategies

Let $G = (X, Y, L)$ be a game and $\Gamma = (E, H, L)$ be the expectation of this game.

Assume that player I employs a mixed strategy $\xi \in E$. This player would like to know his guaranteed gain, i.e. the least gain he can secure for sure even if player II employs his best mixed strategy η. The situation with player II using a pure strategy should not be excluded too.

Denote the guaranteed gain of player I with the strategy ξ by $A_g(\xi)$. Evidently, this is the lower bound of the payoff function $L(\xi, \eta)$ for a given ξ and various $\eta \in H$, i.e.

$$A_g(\xi) = \min_{\eta} L(\xi, \eta) \tag{9.21}$$

At this point player I would like to choose from the set of possible strategies $\xi \in E$ such that guarantees a maximum winning, in other words, he will be interested in the upper bound of the function $A_g(\xi)$, which we shall denote by α_g and call the *minorant value of the game* in mixed strategies,

$$\alpha_g = \max_{\xi} A_g(\xi) = \max_{\xi} \min_{\eta} L(\xi, \eta) \tag{9.22}$$

A strategy satisfying the condition (9.22) is called the *maximin strategy* of player I and denoted by ξ_0.

Assume now that player II has chosen a mixed strategy $\eta \in H$. He would like to know what its maximum loss will be at the best strategy of I, i.e. the upper bound of $L(\xi, \eta)$ at a given η for various $\xi \in E$, denoted by

$$B_g(\eta) = \max_{\xi} L(\xi, \eta) \tag{9.23}$$

Player II is interested in a strategy $\eta \in H$ such that minimizes his loss, i.e. in the lower bound of the function $B_g(\eta)$, that will be denoted by β_g and called the *majorant value of the game* in mixed strategies

$$\beta_g = \min_{\eta} B_g(\eta) = \min_{\eta} \max_{\xi} L(\xi, \eta) \tag{9.24}$$

A strategy satisfying the condition (9.24) is called the *minimax strategy* of player II and denoted by η_0.

For convenience we introduce also the designations for the minorant and majorant values of the game $G = (X, Y, L)$ in pure strategies as follows.

$A_G(x) = \min_y (x, y)$ is the guaranteed gain of I at $x \in X$;
$\alpha_G = \max_x A_G(x)$ is the lower bound of G in pure strategies;
$B_G(y) = \max_x L(x, y)$ is the largest payoff of II at $y \in Y$;
$\beta_G = \min_y B_G(y)$ is the minimax of G in pure strategies.

The rationale behind the mixed strategies follows from the theoren given below.

Theorem 9.2. Given a game $G = (X, Y, L)$ and its averaging $\Gamma = (E, H, L)$, then

$$\begin{aligned}
&\text{(i)}\ A_g(\xi) = \min_y L(\xi, y) \\
&\text{(ii)}\ \alpha_G \leqslant \alpha_g \\
&\text{(iii)}\ B_g(\eta) = \max_x L(x, \eta) \\
&\text{(iv)}\ \beta_G \geqslant \beta_g \\
&\text{(v)}\ \alpha_g \leqslant \beta_g
\end{aligned} \tag{9.25}$$

Item (i) states that if I chooses any mixed strategy, his guaranteed gain equals the guaranteed gain with II employing pure strategies only. According to (ii) the maximin in mixed strategies of I is not inferior to the maximin in pure strategies, i.e. there exists a mixed strategy which is not worse than the optimal pure strategy. Items (iii) and (iv) are similar assertions about player II. Item (v) states that the maximin in mixed strategies α_g does not exceed the minimax of

the game, i.e. with the best mixed strategies the guaranteed gain of I will not exceed the guaranteed loss of II.

Proof. (i) Recognizing that pure strategies $y \in Y$ are a particular case of mixed strategies $\eta \in H$, then $L(\xi, y)$ enters as a particular case in $L(\xi, \eta)$, i.e. is a subset of $L(\xi, \eta)$. By the theorem on the lower bound of a subset (see Sect. 2.1) we have

$$\min_y L(\xi, y) \geqslant \min_\eta L(\xi, \eta) = A_g(\xi) \tag{9.26}$$

$L(\xi, \eta)$ may be viewed as the average of $L(\xi, y)$ taken over all y with the probability distribution $\eta(y)$. By the theorem on the mean (see Sect 5.4) we have

$$\min_y L(\xi, y) \leqslant L(\xi, \eta) \tag{9.27}$$

This inequality holds for any $\eta \in H$ including η_0 for which $L(\xi, \eta_0) = \min_\eta L(\xi, \eta)$.

Therefore

$$\min_y L(\xi, y) \leqslant \min_\eta L(\xi, \eta) = A_g(\xi) \tag{9.28}$$

Comparing (9.26) and (9.28) proves (i).

(ii) Comparing $A_G(x) = \min_y L(x, y)$ with $A_g(\xi)$ which according to (i) may be defined in the same manner as $A_G(x)$ with pure strategies of II, and in view of the fact that a pure strategy is a particular case of a mixed strategy $\xi(x)$ we see that $A_G(x)$ is a subset of $A_g(\xi)$. By the theorem of the upper boundary of a subset we have

$$\max_x A_G(x) \leqslant \max_\xi A_g(\xi)$$

which proves item (ii).

Items (iii) and (iv) can be proved in a similar way.

(v) By definition

$$A_g(\xi) = \min_\eta L(\xi, \eta) \leqslant L(\xi, \eta)$$

$$B_g(\eta) = \max_\xi L(\xi, \eta) \geqslant L(\xi, \eta)$$

Hence,

$$A_g(\xi) \leqslant L(\xi, \eta) \leqslant B_g(\eta)$$

This relationship holds for any ξ and η, i.e. for a ξ such that $A_g(\xi) = \alpha_g$, and for an η such that $B_g(\eta) = \beta_g$, which proves item (v).

Corollary. Comparison of items (ii), (iv) and (v) yields

$$\alpha_G \leqslant \alpha_g \leqslant \beta_g \leqslant \beta_G \tag{9.29}$$

From (9.29) it follows that if G has a saddle point, i.e. $\alpha_G = \beta_G = V$, then $\alpha_g = \beta_g = V$, i.e. I also has a value. Therefore, an optimal

strategy in G is also optimal in Γ. In this case Γ is not worth considering, and optimal strategies may be evaluated by the method described for strictly determined games.

Recalling that game Γ has been brought into the pattern for the purpose of analysis of games with $\alpha_G < \beta_G$, we are faced with a question of whether or not in this case $\alpha_g = \beta_g = V_g$ and V_g may be viewed for a large class of games as the value of game Γ. We prove that this takes place indeed, so that mixed strategies evaluate the value and optimal strategies in this case as well. The definition of an optimal strategy in the circumstances will be as follows.

Given a game $G = (X, Y, L)$ with $\Gamma = (E, H, L)$. If Γ has the value $V_g = \alpha_g = \beta_g$, then we say that G has the value V_g, and any optimal strategy of Γ, i.e. a strategy that guarantees the gain for I and loss to II equal to V_g, is called an *optimal* strategy in G. It is not hard to see that if $\alpha_g = \beta_g = V_g$, then the maximin and minimax strategies $\xi_0(x)$ and $\eta_0(y)$ will be optimal strategies.

The theorem stating that any finite game has a value and each player has optimal strategies is a fundamental theorem of the theory of games and will be proved in the ensuing section.

9.3 The Fundamental Theorem of Game Theory

9.3.1 The *S*-Game

Games in which player I has a finite set of pure strategies may be given a useful geometrical interpretation. Let $G = (X, Y, L)$ be a game with the matrix $\mathbf{Q} = [q_{ij}]$ defined by (9.7). We associate with each $y \in Y$ a point C in m-dimensional space spanned on the coordinates that are payoffs of II at all possible pure strategies of I, i.e.

$$C = [L(x_1, y), \ldots, L(x_m, y)] \tag{9.30}$$

Recognizing that $L(x_i, y_j) = q_{ij}$ and giving y all possible values, we obtain n points

$$C_1 = (q_{11}, \ldots, q_{m1}); \ \ldots; \ C_n = (q_{1n}, \ldots, q_{mn}) \tag{9.31}$$

whose coordinates are the losses designated by the respective columns of the payoff matrix. We shall call a game specified by the set

$$\{C_1, \ldots, C_n\} \tag{9.32}$$

an *S-game*. It is developed as follows.

Player II chooses a point C_i from the set (9.32), say C_3. Independently of this choice, player I chooses a coordinate of point C_i, say the second. The payoff of II will then be equal to the second coordinate of point C_3, i.e. to q_{23}.

As can be seen, the *S*-game is equivalent to a common matrix game, at least when pure strategies are used, as the choice of a point

from the set (9.32) is equivalent to choosing a strategy $y \in Y$, and the choice of the coordinate of this point is equivalent to choosing a strategy $x \in X$.

If the number of strategies available to I is two, then the S-game has a plane representation, as the elements of (9.32) are points of a plane.

Table 9.4

	y_1	y_2	y_3	y_4	y_5
x_1	1	2	−1	0	1
x_2	0	3	1	−1	2

Example 9.2. Consider a game whose matrix is designated by Table 9.4. The equivalent S-game contains five points: $C_1 = (1, 0)$, $C_2 = (2, 3)$, $C_3 = (-1, 1)$, $C_4 = (0, -1)$, and $C_5 = (1, 2)$. The geometrical representation of this game is given in Fig. 9.2.

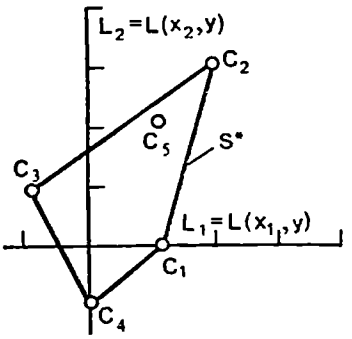


Fig. 9.2. Geometric representation of an S-game

Denote by S^* the convex hull of the finite set $\{C_1, \ldots, C_n\}$ and prove a theorem that states that an S-game is equivalent to a usual matrix game not only in pure but also in mixed strategies.

Theorem 9.3. Any mixed strategy of player II may be represented by a point that belongs to a convex hull S^*, and conversely, any point $S \in S^*$ may be viewed as a mixed strategy of player II.

Proof. Consider mixed strategies $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$ and $\eta = (\eta^{(1)}, \ldots, \eta^{(n)})$ which lead to the payoffs of II equal to

$$L(\xi, \eta) = \sum_i \sum_k q_{ik}\xi^{(i)}\eta^{(k)} = \sum_i \xi^{(i)} \sum_k q_{ik}\eta^{(k)} = \sum_k \xi^{(i)}S^{(i)} \qquad (9.33)$$

where

$$S^{(i)} = \sum_k q_{ik}\eta^{(k)} \qquad (9.34)$$

Denote by S a point in m-dimensional space with the coordinates $S^{(i)}$ where in view of (9.34)

$$\begin{aligned} S^{(1)} &= q_{11}\eta^{(1)} + \ldots + q_{1n}\eta^{(n)} \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ S^{(m)} &= q_{m1}\eta^{(1)} + \ldots + q_{mn}\eta^{(n)} \end{aligned} \qquad (9.35)$$

Recalling that $(q_{1i}, \ldots, q_{mi}) = C_i$, the above system rewrites as

$$S = C_1\eta^{(1)} + \ldots + C_n\eta^{(n)} = \sum_h C_i\eta^{(i)} \tag{9.36}$$

If we observe that the quantities $\eta^{(i)}$ satisfy all the relations (5.3), we see that S is but the weighted average of $C_1, \ldots, C_n$ with the weightings $\eta^{(1)}, \ldots, \eta^{(n)}$, consequently, S is a point that belongs to the convex hull S^*. Therefore, each strategy $\eta = (\eta^{(1)}, \ldots, \eta^{(n)})$ of II will be associated with some point belonging to the convex hull S^*, and specifying this point is equivalent to specifying a mixed strategy of player II.

The converse is also true. Since any point S belonging to the convex hull S^* may be represented as a weighted average of the points $C_1, \ldots, C_n$ specifying the convex hull S^*, i.e. may be represented by the sum (9.36), then for every point $S \in S^*$ there exist weights $\eta^{(1)}, \ldots, \eta^{(n)}$ which specify a mixed strategy of player II.

Corollary. Since the mixed strategy of I remains in the S-game the same as in the original matrix game, from the above theorem it follows that the S-game is completely equivalent to the usual game, i.e. any game may be represented as an equivalent S-game.

In what follows we shall denote the S-game by Γ_1. In going from game $\Gamma = (E, H, L)$ to the S-game the set of mixed strategies of player II, $H = \{\eta_1, \eta_2, \ldots\}$ must give way to the set of S-strategies, i.e. the convex hull S^*. If we denote the payoffs of player II in the S-game by L_1, then the S-game depends on E, S^* and L_1, where the payoffs L_1 are evaluated on the direct product $E \times S^*$. Therefore, the expression

$$\Gamma_1 = (E, S^*, L_1) \tag{9.37}$$

defines an S-game. In view of (9.33) the payoff function is

$$L_1(\xi, S) = \sum_i \xi^{(i)}S^{(i)} = \xi S \tag{9.38}$$

where ξS denotes the dot product of the vectors $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$ and $S = (S^{(1)}, \ldots, S^{(m)})$.

9.3.2 Maximin and Minimax in the S-Game

If player I employs a mixed strategy $\xi \in E$ in the S-game, then his guaranteed gain will be

$$A_{g1}(\xi) = \min_S L_1(\xi, S) = \min_S \xi S \tag{9.39}$$

Denote by $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(m)})$ a strategy of I such that maximizes $A_{g1}(\xi)$. The value of $A_{g1}(\xi)$ will be equal to the maximin of the game, α_{g1}, which owing to the equivalence of the S-game to the

usual matrix game coincides with α_g. Consequently,

$$\alpha_g = A_{g1}(\xi_0) = \max_{\xi} A_{g1}(\xi) = \max_{\xi} \min_{S} \xi S \tag{9.40}$$

A strategy ξ_0 satisfying (9.40) is called the *maximin strategy* of player I.

Assume now that player II employs a strategy $S \in S^*$ so that his payoff is

$$B_{g1}(S) = \max_{\xi} L_1(\xi, S) = \max_{\xi} \xi S \tag{9.41}$$

Define $S_0 = (S_0^{(1)}, \ldots, S_0^{(m)})$ to be a point of the convex hull S^* that minimizes $B_{g1}(S)$. The value of $B_{g1}(S_0)$ will be equal to the minimax of the game β_{g1} which, owing to the equivalence of the S-game to the usual matrix game, coincides with β_{g1}. Hence,

$$\beta_g = B_{g1}(S_0) = \min_{S} B_{g1}(S) = \min_{S} \max_{\xi} \xi S \tag{9.42}$$

The strategy S_0 satisfying (9.42) is called the *minimax strategy* of player II.

The expressions for $B_{g1}(S)$ and β_g may be represented in a more convenient form with the help of the following theorem.

Theorem 9.4. If S is an arbitrary point of m-dimensional space and $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$ is a multivariate whose components satisfy (5.3), then

$$\max_{\xi} \xi S = \max(S^{(1)}, \ldots, S^{(m)}) \tag{9.43}$$

Proof. Let $S^{(k)} = \max(S^{(1)}, \ldots, S^{(m)})$. Consider a particular value of ξ, corresponding to the case

$$\begin{aligned} \xi^{(i)} &= 1, \quad \text{for} \quad i = k \\ &= 0, \quad \text{for} \quad i \neq k \end{aligned} \tag{9.44}$$

In this case $\xi S = S^{(k)}$. Thus, $S^{(k)}$ is a particular value of the scalar product ξS, and hence, a subset of the set ξS obtained for all possible ξ. By the theorem of the boundary of a subset we find

$$S^{(k)} = \max(S^{(1)}, \ldots, S^{(m)}) \leqslant \max_{\xi} \xi S \tag{9.45}$$

On the other hand, substituting the maximum value $S^{(k)}$ for $S^{(1)}, \ldots, S^{(m)}$ in the expression for ξS, we get

$$\xi S = \sum_i \xi^i S^i \leqslant S^{(k)} \sum_i \xi^{(i)} = S^{(k)} \tag{9.46}$$

This expression is valid for any ξ satisfying (5.3) including the one maximizing ξS. Comparing (9.45) and (9.46), we arrive at (9.43).

Utilizing this theorem, the expression for $B_{g1}(S)$ can be written as

$$B_{g1}(S) = \max_{\xi} \xi S = \max(S^{(1)}, \ldots, S^{(m)}) \tag{9.47}$$

Two corollaries follow from this expression.

Corollary 1. From $B_g \leqslant B_{g1}(S)$ we have

$$\beta_g \leqslant \max(S^{(1)}, \ldots, S^{(m)}) \tag{9.48}$$

Any point $S \in S^*$ has at least one coordinate not smaller (i.e. greater than or equal to) than B_g.

Corollary 2. Letting $S = S_0 = (S_0^{(1)}, \ldots, S_0^{(m)})$ we have

$$\beta_g = B_{g1}(S_0) = \max(S^{(1)}, \ldots, S_0^{(m)}) \tag{9.49}$$

The majorant value of the game, B_g, is equal to the greatest coordinate of point S_0 defining the minimax strategy of player II.

9.3.3 The Minimax Theorem

The possibility for each player to evaluate the best strategy is based on the following theorem that may be treated as a proof of existence of solutions for finite games.

Theorem 9.5. Every finite game has a value and each player has at least one optimal strategy.

Outline. Let $G = (X, Y, L)$ be a finite game and $\Gamma = (E, H, L)$ the averaging of this game. It will be convenient to prove the theorem in the S-game setting, therefore, we denote the equivalent S-game by $\Gamma_1 = (E, S^*, L_1)$.

The maximin and minimax for this game will be respectively α_g and β_g, independently of whether the Γ game or the equivalent S-game, Γ_1, is considered, and as indicated earlier $\alpha_g \leqslant \beta_g$.

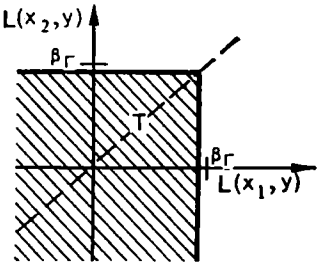


Fig. 9.3. Two-dimensional region T

In order to prove the theorem it will be sufficient to demonstrate that $\beta_g \leqslant \alpha_g$, since a comparison with the preceding inequality yields $\alpha_g = \beta_g$, i.e. the game has a value. To prove this inequality it will suffice to find a mixed strategy ξ_0 of I for which at all $S \in S^*$

$$\beta_g \leqslant \xi_0 S, \quad S \in S^* \tag{9.50}$$

Indeed, if (9.50) holds, then

$$\beta_g \leqslant \min_S \xi_0 S = A_g(\xi_0) = \alpha_g$$

Therefore, the proof of the theorem reduces to the proof of (9.50).

Proof. Consider a set T consisting of all points $t = (t^{(1)}, \ldots, t^{(m)})$ such that $t^{(i)} < \beta_g$ for $i = 0, 1, \ldots, m$. Figure 9.3 shows the re-

gion T in two dimensions, which looks like a rectangular area with the vertex on the 45-degree line. Let us look at the properties of the set T.

The set T is convex. Let t_1 and t_2 be arbitrary points of this set. The equation of the line connecting these points is

$$t = w_1 t_1 + w_2 t_2, \quad w_1, w_2 \geqslant 0, \quad w_1 + w_2 = 1$$

The projection of this equation on the ith axis is in view of Theorem 9.4

$$t^{(i)} = w_1 t_1^{(i)} + w_2 t_2^{(i)} \leqslant \max\,(t_1^{(i)}, t_2^{(i)}) < \beta_g \tag{9.51}$$

Consequently, any point of this line segment belongs to T and the set T is convex.

The set T does not intersect with the set S^*. The property follows from the fact that any point of S^* has at least one coordinate greater or equal to β_g (see Corollary 1 of Theorem 9.4), and hence T and S^* have no points in common.

Because T and S^* are convex nonintersecting regions, there exists a hyperplane separating them and such that T and S^* find themselves in different subspaces defined by this hyperplane. Consequently, there exist $a = (a^{(1)}, \ldots, a^{(m)})$ and a number c such that the equation

$$ax = c \tag{9.52}$$

will be the equation of separating hyperplane, and

$$aS \geqslant c \quad \text{for} \quad S \in S^*, \quad at \leqslant c, \quad \text{for} \quad t \in T \tag{9.53}$$

We demonstrate that $a^{(i)} \geqslant 0$ for $i = 1, \ldots, m$. Let $\delta_i = (\delta_i^{(1)}, \ldots, \delta_i^{(m)})$ be a point whose coordinate i is equal to 1, and other coordinates are a small quantity $\varepsilon > 0$. Consider the point $S_0 \in S^*$. Since its maximum coordinate is β_g (see Corollary 2 of Theorem 9.4), then point $S_0 - \delta_i \in T$. Consequently,

$$aS_0 \geqslant c \geqslant a\,(S_0 - \delta_i)$$

whence it follows that

$$a\delta_i = a^{(1)}\,\delta_i^{(1)} + \ldots + a^{(m)}\,\delta_i^{(m)} \geqslant 0$$

If $\varepsilon \to 0$, then $\delta_i^{(k)} \to 0$ at $k \neq i$ and $\delta_i^{(i)} = 1$. The last condition yields

$$a^{(i)} \geqslant 0, \quad i = 1, \ldots, m \tag{9.54}$$

Denote

$$\xi_0 = a / \sum_i a^{(i)} \tag{9.55}$$

Evidently $\xi_0 \in E$ since

$$\xi_0^{(i)} = a^{(i)} / \sum_i a^{(i)} \geqslant 0, \quad \sum_i \xi_0^{(i)} = 1$$

Denote also

$$v = c/\sum_i a^{(i)} \tag{9.56}$$

Dividing (9.53) by $\sum_i a^{(i)}$ and observing (9.55) and (9.56), we have

$$\begin{aligned} \xi_0 S &\geqslant v \quad \text{for} \quad S \in S^* \\ \xi_0 t &\leqslant v \quad \text{for} \quad t \in T \end{aligned} \tag{9.57}$$

Let t_0 be a point with the coordinates $t_0^{(i)} = \beta_g - \varepsilon$, $\varepsilon > 0$, for $i = 1, \ldots, m$. Clearly, $t_0 \in T$. From the second inequality of (9.57) we obtain

$$\xi_0 t_0 \leqslant v \tag{9.58}$$

Let $\varepsilon \to 0$ so that $t_0^{(i)} \to \beta_g$, then

$$\xi_0 t_0 = \sum_i \xi_0^{(i)} t_0^i \to \beta_g \sum_i \xi_0^{(i)} = \beta_g \tag{9.59}$$

Comparing (9.58) and (9.59), we find

$$v \geqslant \beta_g \tag{9.60}$$

The first inequality of (9.57) gives

$$\xi_0 S \geqslant v \geqslant \beta_g \tag{9.61}$$

which proves (9.50).

Thus, $v = \alpha_g = \beta_g$ is a value of the game and ξ_0 and S_0 are optimal mixed strategies of the players.

9.3.4 Geometric Interpretation of the Minimax Principle

Here we wish to demonstrate that the point S_0 that defines a minimax strategy of player II is a boundary point of S^* and such where T is tangent to S^*.

As indicated $S_0 = (S_0^{(1)}, \ldots, S_0^{(m)}) \in S^*$ and $\max (S_0^{(1)}, \ldots, S_0^{(m)}) = \beta_g$. On the other hand, $t = (t^{(1)}, \ldots, t^{(m)}) \in T$ if $t^i < \beta_g$ for $i = 1, \ldots, m$.

Consider a point $S_0' = (S_0^{(1)} - \varepsilon, \ldots, S_0^{(m)} - \varepsilon)$ with $\varepsilon > 0$. Obviously, $\max (S_0^{(1)} - \varepsilon, \ldots, S_0^{(m)} - \varepsilon) < \beta_g$ so that $S_0^{(i)} - \varepsilon < \beta_g$, for $i = 1, \ldots, m$. Consequently, $S_0' \in T$. However, $\lim_{\varepsilon \to 0} S_0' = S_0 \in S^*$. Consequently, (1) S_0 is a boundary point of S^*, (2) S_0 is a point where T is tangent to S^*.

These properties indicate a straightforward way to a minimax strategy S_0 for the case of player I having two pure strategies, i.e. when the equivalent S-game is represented by a set of points on the plane. To construct the region T tangent to S^* it is convenient to

draw a straight line through the origin at 45° with the abscissa bearing the vertex of the rectangular area (called quantant) forming T.

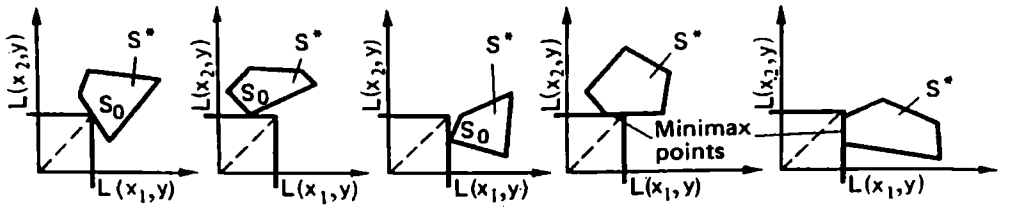


Fig. 9.4. Geometric representation of the minimax strategy

Figure 9.4 depicts various situations of mutual arrangement of S^* and T showing points representing a minimax strategy S_0 for player II.

9.4 General Techniques for Solving Games

9.4.1 Dominant and Admissible Strategies

Mixed strategies are randomized pure strategies $x \in X$ and $y \in Y$ that are taken in agreement with the probability distributions $\xi(x)$ and $\eta(y)$. However, in many situations, some pure strategies are obviously undesirable and in evaluating an optimal mixed strategy they simply need not be taken into account. We call the pure strategies that enter an optimal mixed strategy *admissible*. To facilitate the evaluation of admissible strategies we introduce the notion of *dominant* strategies.

Consider two strategies y_l and y_p for player II. Let I employ a strategy x_i. Then II loses respectively q_{il} and q_{ip}. It may so happen that for any i

$$q_{il} \leqslant q_{ip}, \quad i = 1, \ldots, m \tag{9.62}$$

that is, the entries in column l of the loss matrix do not exceed the respective entries in column p. This implies that y_p is not advantageous for II under any conditions as this strategy entails greater losses than y_l. Therefore, y_p should be discarded from the matrix of the game. The strategy y_l satisfying (9.62) is said to be dominant with respect to y_p.

Dominant strategies of II receive a geometric interpretation in going to the equivalent S-game on the plane. In this case $m = 2$ and (9.62) can be rewritten as

$$q_{1l} \leqslant q_{1p}, \quad q_{2l} \leqslant q_{2p} \tag{9.63}$$

Figure 9.5 graphs two placements of points C_l and C_p corresponding to pure strategies y_l and y_p. It is obvious that at (*a*) strategy C_l

dominates C_p, while at (b) none of the strategies is dominant. In order that strategy y_l dominate y_p, point C_l must be on the left and below C_p.

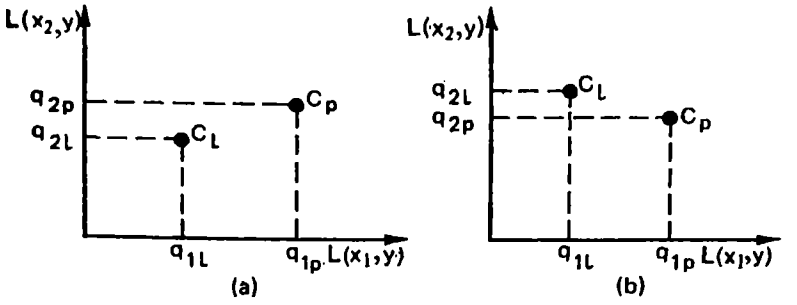


Fig. 9.5. Evaluation of dominant strategy in an S-game

Dominant strategies of player I may be determined in a similar manner. We say that strategy x_l dominates x_p if the gain of I for x_l is larger than gains with x_p for any strategy $y \in Y$, i.e.

$$q_{li} \geqslant q_{pi}, \quad i = 1, \ldots, n \tag{9.64}$$

In other words, entries in row x_l of the loss matrix are greater than the respective losses in row x_p.

Example 9.3. Figure 9.6 specifies an S-game by points C_1 through C_6 on the plane. Obviously, C_1 dominates C_5, while C_1, C_2, C_6 dominate C_4. Discarding C_5 and C_4, we arrive at an S-game specified by points C_1, C_2, C_3 and C_6 none of which dominates the others.

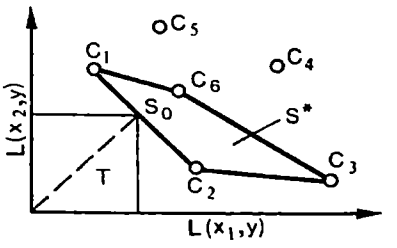


Fig. 9.6. Example of an S-game

Elimination of the dominant strategies from the matrix simplifies the game appreciably, and facilitates the search for an optimal strategy. Assume that in the game under study no one strategy dominates another. It would be interesting to answer then whether in this case all strategies are admissible, i.e. may be used in the construction of an optimal mixed strategy.

Let us again turn to the game specified in Fig. 9.6. Once points C_4 and C_5 have been deleted as dominated, the remaining points are C_1, C_2, C_3 and C_6. Constructing the S_0-lower quantant T we see that point S_0 defining the strategy for II lies on the line connecting C_1 and C_2 and may be represented as a weighted average of these points. Therefore, an optimal mixed strategy for II is a combination of pure strategies y_1 and y_2 which in this case are admissible. Stra-

tegies y_3 and y_6 are not useful any longer and their use would be undesirable.

The total number of admissible strategies for each player may be determined by observing that the convex hull S^* of the finite set $\{C_1, \ldots, C_n\}$ is a convex polyhedron in m-dimensional space. Point S_0 is a boundary point for the convex polyhedron S^* and lies on one of its faces whose vertices correspond to the admissible strategies for player II. If we recognize that the number of vertices of any face of S^* cannot exceed the total number of vertices in the polyhedron, i.e. n, and cannot exceed the dimensionality of the space, i.e. m, we arrive at a conclusion that the number of strategies of II does not exceed the least of the numbers m and n. Because the assignments of I and II to either player are conditional, a similar conclusion will be true for I.

Thus, in an $m \times n$ matrix game the number of admissible strategies for each player does not exceed the least of the numbers m and n.

Theorem 9.6. If one of the players sticks to his optimal mixed strategy, then the gain of players remains unchanged and equal to the value of the game, v, independently of mixed (or pure) strategy the opponent employs within the limits of his admissible strategies.

Proof. Let v be the value of the game, $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(m)})$ be the optimal mixed strategy of player I. If I employs his optimal strategy ξ_0, then his gain cannot be smaller than v.

Assume that player II has k admissible strategies denoted $y_1, \ldots, y_k$. If I chooses his optimal strategy ξ_0, and II his useful strategy y_i $(i = 1, \ldots, k)$, then the gain for I, v_i, will not be smaller than v, i.e.

$$v_i \geqslant v, \quad i = 1, \ldots, k \tag{9.65}$$

Consider now an optimal mixed strategy of player II, η_0. It is specified by the probabilities $\eta_0^{(i)}$ which weigh the pure admissible strategies involved. It is essential that for $i = 1, \ldots, k$, $\eta_0^{(i)}$ cannot be all zeros, as in this case strategy j would not enter the optimal mixed strategy of player II, that is, would not be admissible.

Let us determine the expected gain for I if II employs his optimal mixed strategy. This setting means that I will obtain gains $v_1, \ldots, v_k$ corresponding to pure strategies $y_1, \ldots, y_k$ with probabilities $\eta_0^{(1)}, \ldots, \eta_0^{(k)}$. The expected gain is then

$$I_{\exp} = v_1\eta_0^{(1)} + \ldots + v_k\eta_0^{(k)} = \sum_{i=1}^{k} v_i\eta_0^{(i)} \tag{9.66}$$

Replacing v_i with v and using (9.65), we get

$$I_{\exp} \leqslant v \sum_{i=1}^{k} \eta_0^{(i)} = v \tag{9.67}$$

At optimal strategies, the expected gain is, however, the value of the game, v, i.e.

$$L_{\text{exp}} = v \tag{9.68}$$

On the other hand, (9.66) can become (9.68) only subject to the condition

$$v_i = v, \quad i = 1, \ldots, k \tag{9.69}$$

which proves the theorem.

Therefore, the expected gain of I using his optimal mixed strategy will be the same if II employs any of his admissible strategies, and, hence, any mixed strategy that consists of the admissible strategies.

It should be kept in mind, of course, that optimal mixed strategies are more advantageous under any circumstances. Useful strategies, although nonoptimal, will be harmless until the opponent adheres to his optimal strategy. The use of an optimal strategy always provides a gain equal to the value of the game.

Now, it remains to clarify how admissible and optimal strategies can be evaluated, i.e. how the game can be solved.

9.4.2 Evaluation of Optimal Strategies

Consider a game $G = (X, Y, L)$ with an $m \times n$ matrix. We shall assume that G has no saddle point — otherwise a solution can be easily found by the rule established in Sect. 9.2 — and all dominated strategies have been deleted.

Denote by $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(m)})$ an optimal mixed strategy for player I. Some $\xi_0^{(i)}$ may be zeros, which implies that the respective strategy x_i is no longer admissible. We seek $\xi_0^{(i)}$ for $i = 1, \ldots, m$.

Suppose that II employs strategy y_k. If this is an admissible strategy, then the gain of I will be v, if it is inadmissible, then the gain of I may be greater than v. Hence, in the general case,

$$L(\xi_0, y_k) = \xi_0^{(1)} q_{1k} + \ldots + \xi_0^{(m)} q_{mk} \geqslant v \tag{9.70}$$

Expressions of this type may be compiled for each pure strategy of II, i.e. for $k = 1, \ldots, n$. In addition, it is known that

$$\xi_0^{(i)} \geqslant 0, \quad \xi_0^{(1)} + \ldots + \xi_0^{(m)} = 1 \tag{9.71}$$

Let $v > 0$; this holds when all elements of matrix q_{ik} are positive. If some of them are negative, they can be made positive by adding to all elements of the matrix some number $v' > 0$. Then the value of the game increases by v'.

Divide (9.70) and (9.71) by v and denote

$$\xi_0^{(i)}/v = p_i \tag{9.72}$$

It should be evident that $p_i \geqslant 0$. The inequalities (9.70) for $k = 1, \ldots, n$ can now be written as

$$\begin{aligned} p_1 q_{11} + \ldots + p_m q_{m1} &\geqslant 1 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ p_1 q_{1n} + \ldots + p_m q_{mn} &\geqslant 1 \end{aligned} \tag{9.73}$$

and (9.71) takes the form

$$p_1 + \ldots + p_m = 1/v \tag{9.74}$$

Here, $p_1, \ldots, p_m$ are unknown quantities that define the mixed strategy of player I. In order to arrive at an optimal mixed strategy we need a solution to (9.73) such that maximizes v, and minimizes the objective function (9.74). Taken together, this is a common LP problem.

For technical convenience the inequalities in (9.73) should be made equations by subtracting from the right-hand sides some positive additional variables z_i, namely,

$$p_1 q_{1i} + \ldots + p_m q_{mi} - z_i = 1, \quad i = 1, \ldots, n \tag{9.75}$$

In those equations of (9.75) that correspond to admissible strategies for II, the solution yields $z_i = 0$. Thus, solving the system (9.75) subject to the minimization of the objective function (9.74), we obtain a mixed strategy for I, the value of the game v, and admissible strategy of player II, $y_1, \ldots, y_k$.

To determine an optimal mixed strategy for II, $\eta_0 = (\eta_0^{(1)}, \ldots, \eta_0^{(k)})$ involving k unknowns $\eta_0^{(1)}, \ldots, \eta_0^{(k)}$ we need to compile k equations. One of them is

$$\eta_0^{(1)} + \ldots + \eta_0^{(k)} = 1 \tag{9.76}$$

The other $k - 1$ equations can be derived by writing the expressions for expected losses for an optimal mixed strategy of II and for any $k - 1$ useful strategies of I. For example, for a useful strategy x_i we have

$$\eta_0^{(1)} q_{i1} + \ldots + \eta_0^{(k)} q_{ik} = v \tag{9.77}$$

Having derived $k - 1$ equations of the type (9.77) augmented by (9.76), we obtain k equations that can be solved to find $\eta_0^{(1)}, \ldots, \eta_0^{(k)}$ yielding an optimal mixed strategy of player II.

Example 9.4. Solve a game specified by the left part of Table 9.5. First of all we should check that the game has no saddle points and

Table 9.5

	y_1	y_2	y_3	y_1	y_2	y_3
x_1	2	−3	4	7	2	9
x_2	−3	4	−5	2	9	0
x_3	4	−5	6	9	0	11

none of the strategies dominates the other. To make all the entries of the matrix positive, we add 5 to each entry, then the matrix takes the form shown in the right-hand side of the table.

Equations (9.75) will be written as

$$\begin{aligned} 7p_1 + 2p_2 + 9p_3 - z_1 &= 1 \\ 2p_1 + 9p_2 \qquad - z_2 &= 1 \\ 9p_1 \qquad + 11p_3 - z_3 &= 1 \end{aligned}$$

We seek to solve these equations subject to

$$p_1 + p_2 + p_3 = \min$$

Solving this LP problem yields $z_1 = z_2 = z_3 = 0$, $p_1 = 0.05$, $p_2 = 0.1$, and $p_3 = 0.05$.

As can be seen, all three strategies (y_1, y_2, and y_3) are admissible. From (9.74) we find

$$v = 1/(p_1 + p_2 + p_3) = 5$$

with

$$\begin{aligned} \xi_0^{(1)} &= 5p_1 = 0.25 \\ \xi_0^{(2)} &= 5p_2 = 0.5 \\ \xi_0^{(3)} &= 5p_3 = 0.25 \end{aligned}$$

To obtain an optimal mixed strategy for II we set up one equation of the type (9.76), namely,

$$\eta_0^{(1)} + \eta_0^{(2)} + \eta_0^{(3)} = 1$$

and two equations of the type (9.77) for the admissible strategies x_2 and x_3

$$2\eta_0^{(1)} + 9\eta_0^{(2)} = 5, \quad 9\eta_0^{(1)} + 11\eta_0^{(3)} = 5$$

Solving the three equations simultaneously we find $\eta_0^{(1)} = \eta_0^{(3)} = 0.25$, and $\eta_0^{(2)} = 0.5$.

Recalling that 5 was added to all elements of the matrix, we find the value of the game $v - 5 = 0$.

9.4.3 Geometric Solution of $2 \times n$ Matrix Games

In a $2 \times n$ rectangular game, a mixed strategy of player I is an ordered pair $\xi = (\zeta, 1 - \zeta)$. If II employs a strategy y_k, then the expected gain for I is

$$L(\zeta, y_k) = \zeta q_{1k} + (1 - \zeta) q_{2k} = q_{2k} + \zeta (q_{1k} - q_{2k}) \qquad (9.78)$$

which is a linear function of ζ, as shown in Fig. 9.7.

Evaluation of an optimal strategy ξ_0 will be discussed with reference to a matrix game specified by Table 9.6. The geometric representation of this game is given in Fig. 9.8. We draw this diagram with a baseline [0, 1] and indicate on it, by the distance from 0, the relative frequency of the second row in the strategy x_2. The two payoffs for II, by his first and his second strategy, dependent on I's choosing one of his pure strategies, are drawn as ordinates above 0 and 1, respectively. We connect the end points of the ordinates belonging to the same pure strategy of I, and call these lines the strategy lines of I. They are labeled by 1, 2, and 3 for y_1, y_2, and y_3, respectively.

Table 9.6

	y_1	y_2	y_3
x_1	1	3	5
x_2	4	2	1

The guaranteed gain of I at any ζ is emphasized by the bold line. According to the minimax principle, an optimal strategy ξ_0 is the

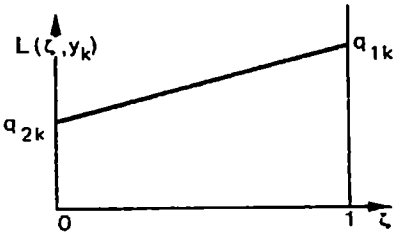


Fig. 9.7. A strategy line for strategy y_k

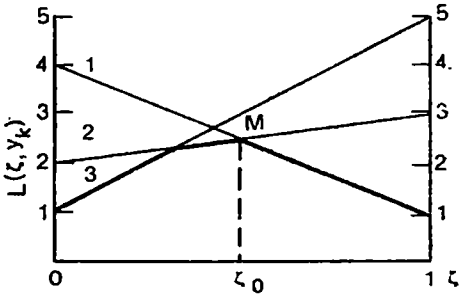


Fig. 9.8. Evaluation of a minimax strategy in a $2 \times m$ game

one at which the guaranteed gain is largest. This strategy is located by point M, the intersection of the strategy lines for y_1 and y_2, which thereby will be the admissible useful strategies of player II.

The optimal mixed strategy ξ_0 may be found by equating the expected gains of I for the admissible strategies y_1 and y_2 of player II,

$$\zeta_0 + 4\,(1 - \zeta_0) = 3\zeta_0 + 2\,(1 - \zeta_0)$$

which yields $\zeta_0 = 1/2$. Thus, $\xi_0 = (1/2, 1/2)$. Now, the value of the game can be obtained as the expected gain of I at the strategy ξ_0 and any of the admissible strategies, say y_1, of II, namely,

$$v = \zeta_0 + 4\,(1 - \zeta_0) = 2.5$$

Given the value of the game v, the optimal mixed strategy of II, $\eta_0 = (\eta_0^{(1)}, 1 - \eta_0^{(1)}, 0)$, can be found from the expected payoffs of II at any of the admissible strategies of I. For example, for x_1 we have

$$\eta_0^{(1)} + 3\,(1 - \eta_0^{(1)}) = 2.5$$

which gives $\eta_0^{(1)} = 1/4$. Thus, $\eta_0 = (1/4, 3/4, 0)$.

EXERCISES

9.1. Compute the loss matrix for the game of Example 9.1.

9.2. Determine the payoff matrix by drawing the game tree for the following game. A symmetric coin is tossed and if it lands heads up, then I chooses 1 or 6, if it comes tails up then 2 or 7. Then player II chooses 3 or 9. The chosen numbers are added. Again a coin is thrown, the probability that it lands heads up being 0.8. If it comes heads up, then the entire sum goes to player I. If it lands tails up, then I pays this sum to II. Player I knows the outcome of the first random move. Player II is unaware of the outcomes of the preceding moves.

9.3. A convenient randomization mechanism is a position of the seconds hand in a clock at a random time instant. Describe how this mechanism can be used to obtain the probability distributions as follows: (3/5, 2/5), (1/6, 2/3, 1/6), and (1/5, 2/5, 2/5).

9.4. Describe the actions of players in one play of the game specified by Table 9.3, given that player I uses the pure strategy x_2, and player II employs a mixed strategy $\eta = (0.5, 0.5)$.

10

STATISTICAL DECISION THEORY

10.1 Statistical Games

10.1.1 Strategic and Statistical Games

Statistical games represent a special model that has a great significance in analysis of various situations of practical life. This model has a number of features that differ it from *strategic games* considered thus far. Strategic games are based on the assumption that the interests of the players are contradictory. Each player tends to choose his strategy such that to receive maximum benefit from the game and reduce the opponent's gain to a minimum. It is assumed also that the players are active adversaries and tend to use optimal strategies as far as possible.

In many practical situations, however, one of the players appears to be neutral, i.e. does not seek for a maximum advantage for himself and, consequently, does not use to his own advantage errors made by the adversary. Games where one of the players is nature relate to this class. By the name "nature" we understand the multitude of external circumstances under which a decision has to be made.

Nature should not be treated as an intelligent opponent, which would benefit from the errors made by man. In other words, nature has no malicious intent with respect to man. It is developing and acting according to its rules and it is up to man to utilize these laws for his own benefit. If man knew the laws of nature exactly, he could use them to maximum advantage. However, in many situations man either does not know the law of nature or has insufficient information about this law.

An unavoidable cost for attempting to make a decision under incomplete information is the likelihood of erroneous decisions. Actual situations are usually such that the decision maker cannot remake a decision. Also he cannot recall a decision as this action is itself a decision and may have undesirable consequences as other decisions do. A way out from this situation is to develop a decision strategy which, although fails to preclude erroneous decisions to be made, minimizes the deleterious consequences of such decisions.

Another possibility is to study the opponent, i.e. nature, by conducting experiments. Theoretically, by carrying out an unlimited experiment we might improve our knowledge of nature as far as de-

sirable to act under complete certainty. However, there are two obstacles to this approach: (1) such an experiment would take time whereas in many cases a fast decision is needed, and (2) such an experiment may be very costly, as much or even more than the gain obtained with the additional knowledge from the experiment.

It should be evident that an important question to answer in a game against nature is whether an experiment must be conducted and, if so, then what kind of experiment, when it should be terminated, and what actions should be taken after it is over.

Games in which one opponent is nature and the other is man are called *statistical games*, and the theory of such games is known as *statistical decision theory* [57, 60, 61]. In what follows we shall call the opponent of nature the *statistician*.

10.1.2 Space of States of Nature

Saying a strategy of nature we shall mean the multitude of external circumstances under which a decision is made. This unknown quantity ϑ which affects the decision process is commonly called the *state of nature*. In the general case, there exists a certain set of states of nature $\Theta = \{\vartheta_1, \ldots, \vartheta_m\}$ which will be called, as in Chapter 6, the state-of-nature space, the elements of this space being pure strategies of nature.

If we knew in advance which of its pure strategies nature would employ in each particular case, then we could make a decision based on a perfect information on the state of nature. As a rule, however, only a list of pure strategies is available. In addition, from experience it is known how frequently nature tends to change one or another pure strategy, i.e. a prior probability distribution $\xi(\vartheta)$ on space of states of nature Θ is known. This prior probability distribution is called a *mixed strategy of nature*.

10.1.3 Loss Function

A statistical decision problem can be viewed as a two-person zero-sum game in which the statistician is player II and nature is player I. The statistician is to make a decision, or take an action, from the set of decisions or actions. Denote the possible actions of the statistician $a_1, \ldots, a_l$. Each of these actions is a *pure strategy* of the statistician. The set $A = \{a_1, \ldots, a_l\}$ is the *strategy space for the statistician*.

The statistician is to estimate his actions. For this purpose he assumes that taking an action a he may incur a loss $L(\vartheta, a)$ that depends on both the action a and the unknown state of nature ϑ. The function $L(\vartheta, a)$, called the *loss function*, must be delineated in advance for all possible combinations of $a \in A$ and $\vartheta \in \Theta$, i.e.,

must be specified on the product $\Theta \times A$. It may also be given analytically, or by analogy with (9.7), as a loss matrix

$$Q = [q_{ij}] = \begin{pmatrix} q_{11} & \cdots & q_{1l} \\ \vdots & & \vdots \\ q_{m1} & \cdots & q_{ml} \end{pmatrix} \tag{10.1}$$

where $q_{ij} = L(\vartheta_i, a_j)$. When the loss function is specified, the statistician may choose actions that are best from the point of view of the available information about the state of nature.

Usually the statistician knows the mixed strategy of nature, i.e. the prior probability distribution $\xi(\vartheta)$ over state-of-nature space, Θ. The expected losses to the statistician taking an action will be

$$L(\xi, a) = E_{\vartheta}[L(\vartheta, a)] = \sum_{\vartheta \in \Theta} L(\vartheta, a)\, \xi(\vartheta) \tag{10.2}$$

The best action for the statistician will be *Bayesian*, a^*, at which the loss is a minimum

$$R^*(\xi) = L(\xi, a^*) = \min_{a \in A} L(\xi, a) \tag{10.3}$$

The statistician need not necessarily confine himself to using only one pure strategy. He may employ a combination of pure strategies randomized according to a certain probability distribution. To obtain such a mixed strategy the statistician should specify a probability distribution $\eta(a) = (\eta^{(1)}, \ldots, \eta^{(m)})$ giving the relative frequencies for his pure strategies $a_1, \ldots, a_m$. In the general case, the statistician has a certain set of mixed strategies $\mathbf{H} = \{\eta_1(a), \ldots, \eta_n(a)\}$ called the *mixed-strategy space of the statistician.*

If the statistician employs a mixed strategy $\eta(a)$ and nature uses a mixed strategy $\xi(\vartheta)$, then the expected loss to the statistician is

$$L(\xi, \eta) = E_{\vartheta, a}[L(\vartheta, a)] = \sum_{\vartheta, a} L(\vartheta, a)\, \xi(\vartheta)\, \eta(a) \tag{10.4}$$

In this case the statistician seeks a strategy $\eta^*(a) \in \mathbf{H}$ such that would minimize his expected loss $L(\xi, \eta^*)$, i.e.

$$L(\xi, \eta^*) = \min_{\eta \in H} (\xi, \eta) \tag{10.5}$$

These examples represent a comparatively simple statistical problem that seeks to determine the best strategy for the statistician on the basis of merely a prior information available on states of nature. The statistician does not attempt to refine his knowledge of the actual state of nature by experimentation. Therefore this type of statistical game may be called a *statistical game without sample information or experiment.*

10.1.4 Examples of Statistical Games

For better insight into the structure and solution techniques of statistical games we consider a few examples which should form an illustrative basis for the fundamental concepts of statistical decision theory.

Example 10.1. *A tool replacement problem.* If a venture continues over an extended time, the equipment becomes worn and the need for replacement becomes necessary. The tools are in a three states: ϑ_1, the equipment is operable and maintenance is all that is needed; ϑ_2, some details are worn and require a substantial repair or replacement; and ϑ_3, the crucial details are worn out and further operation of the tools is no longer reasonable.

Past experience with the operation of similar equipment suggests that after years 20% of the equipment may be found in state ϑ_1, 50% in state ϑ_2, and 30% in state ϑ_3.

The manager is faced with three options: a_1, keep the equipment in operation for another year after a limited repair; a_2, to carry out an overhaul using a repair crew; and a_3, replace the equipment with a new one.

The costs, or losses incurred to the venture by the above options are listed in Table 10.1. The losses involve the cost of repair or replacement and the costs due to the loss of efficiency leading to economic decline. The table lists also the prior probabilities (relative frequencies) of the different states of nature, i.e. the mixed strategy of nature, $\xi(\vartheta)$.

Table 10.1. $L(\vartheta, a)$

(ϑ)	a_1	a_2	a_3
Replacement problem			
ϑ_1	0.2	1	3	5
ϑ_2	0.5	5	2	4
ϑ_3	0.3	7	6	3
Processing problem			
ϑ_1	0.6	0	1	3
ϑ_2	0.4	5	3	2

For a given mixed strategy $\xi(\vartheta)$, the expected losses for the different actions are as follows:

$$L(\xi, a_1) = \sum_{\vartheta} L(\vartheta, a_1)\,\xi(\vartheta) = 1 \times 0.2 + 5 \times 0.5 + 7 \times 0.3 = 4.8;$$

$$L(\xi, a_2) = 3.4; \quad L(\xi, a_3) = 3.9$$

Example 10.2. A processing line may receive raw materials with low and high content of impurities, denoted respectively ϑ_1 and ϑ_2. On the average, the supplies have 60% of the first type and 40% of the second type. Three modes of processing, a_1, a_2, and a_3 are used to handle these materials. The prior probabilities of states of nature and losses reflecting the quality of product and expenditures of raw materials depending on its type and the processing mode are summarized in Table 10.1.

The expected losses corresponding to the specified prior probabilities at the different processing modes are as follows:

$$L(\xi, a_1) = 2.0; \quad L(\xi, a_2) = 1.8; \quad L(\xi, a_3) = 2.6$$

The Bayesian action is seen to be a_2.

10.2 Statistical Games Without Sample Information

10.2.1 Reformulation as an S-Game

A statistical game may be represented as an equivalent S-game in essentially the same manner as for strategical games. For this purpose we associate with each pure strategy a_j, $j = 1, \ldots, l$, a point $C_j = (q_{1j}, \ldots, q_{mj})$ of m-dimensional space whose coordinates are the losses to the statistician $L(\vartheta_i, a_j) = q_{ij}$ at the various states of nature ϑ_i, for $i = 1, \ldots, m$. For example, in the problem of a processing line, the pure strategies a_1, a_2, and a_3 of the statistician will be associated with the points $C_1 = (0, 5)$, $C_2 = (1, 3)$, and $C_3 = (3, 2)$ of the plane as shown in Fig. 10.1. The convex hull S^* of the set $\{C_1, C_2, C_3\}$ defines the region of all possible strategies of the statistician, both pure and mixed.

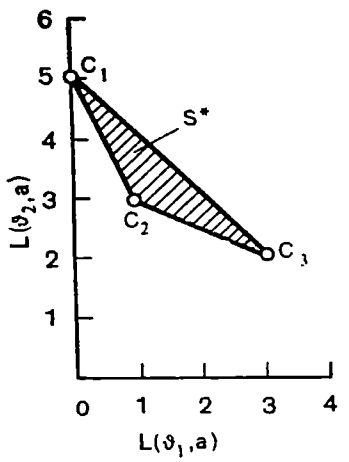


Fig. 10.1. Processing line problems as an S-game

Consider the principles that may provide a guideline for the statistician in choosing his strategy. Statisticians are not consistent in the opinion as to which principle is best in statistical games. To put this in other words, there exists no universal rule that enables one to choose a certain line of behavior independent of the specific situation. However, although there may be disagreements as to what should be done under the circumstances, agreement can be achieved about what should not be done with the aid of the admissibility concept similar to that of strategic games.

10.2.2 Admissible Strategies

Suppose that we are dealing with a mixed strategy of the statistician $\eta(a)$. Two situations are possible.

(1) There exists no strategy better than $\eta(a)$. This means that there exists no strategy $\eta'(a)$ for which

$$L(\vartheta_i, \eta') \leqslant L(\vartheta_i, \eta) \tag{10.6}$$

for all $\vartheta \in \Theta$, although for some ϑ this relation will be true. In this case, the strategy $\eta\ (a)$ is called *admissible*. Such a strategy is not preferable to the others as there may be other strategies of equal interest.

(2) There exists a strategy $\eta'\ (a)$ better than $\eta\ (a)$. This implies that (10.6) holds for $\eta'\ (a)$ for all $\vartheta \in \Theta$. In this case $\eta\ (a)$ is deemed *inadmissible* and dropped from the consideration in favor of $\eta'\ (a)$.

It is convenient to consider admissible strategies in terms of an S-game. Here, a strategy of the statistician is defined by a point S of the convex hull S^* and the losses for various $\vartheta \in \Theta$ are defined by the coordinates of this point. Accordingly, a strategy given by a point S will be admissible if there exists no other point $S' \in S^*$ for which all the coordinates will be smaller than the respective coordinates of S.

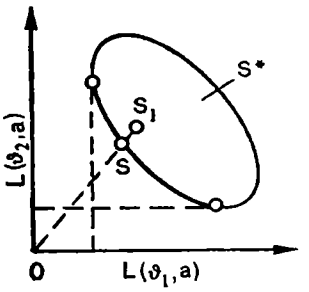


Fig. 10.2. Admissible strategies in S-games

We consider the evaluation of admissible strategies for the case of state-of-nature space consisting of two elements ϑ_1 and ϑ_2. The convex hull S^* corresponding to this situation is graphed in Fig. 10.2.

Consider a strategy, defined by S_1, lying inside S^*. This strategy cannot be deemed admissible as all points lying on OS_1 inside S^* present better strategies than S_1. The best of them is the strategy associated with S that belongs to the bottom lower boundary of S^*. Accordingly, all the internal points may be excluded in favor of points belonging to this section of the boundary of S^* emphasized by heavy line.

Movement along this line does not bring about any advantages, as new points decrease the losses corresponding to one state of nature and increase the losses corresponding to another. Thus, the points belonging to the bottom leftmost boundary of S^* define the space of admissible strategies of the statistician.

Example 10.3. In the processing line problem graphed in Fig. 10.1, the left bottom boundary of S^* consists of the segments C_1C_2 and C_2C_3 each of which is defined by a combination of the pure strategies a_1, a_2 and a_2, a_3. Define $0 \leqslant w \leqslant 1$, then the line segment C_1C_2 may be written as

$$S = wC_1 + (1 - w)\, C_2 \tag{10.7}$$

This equation defines a mixed strategy $\eta\ (a) = (w,\ 1 - w,\ 0)$. Projecting the equation for C_1C_2 onto the coordinate axes, we get

$$\begin{aligned} L\ (\vartheta_1,\ \eta) &= 0 \cdot w + 1 \cdot (1 - w) = 1 - w \\ L\ (\vartheta_2,\ \eta) &= 5 \cdot w + 3 \cdot (1 - w) = 3 + 2w \end{aligned} \tag{10.8}$$

Likewise for the equation of C_2C_3 that defines the mixed strategy $\eta(a) = (0, w, 1-w)$ we write

$$L(\vartheta_1, \eta) = 3 - 2w$$
$$L(\vartheta_2, \eta) = 2 + w \qquad (10.9)$$

10.2.3 Evaluation of Strategies

In different situations the statistician may employ various principles or rules to choose his best mixed strategy. One of possible rules is the minimax principle that is used to advantage in strategical games when the game is performed against an intelligent opponent. However, in a number of situations it is rational to use this principle in statistical games, i.e. assume that nature is an intelligent opponent.

According to the minimax rules the statistician chooses a mixed strategy $\eta(a)$ such that minimizes his expected losses $L(\vartheta, \eta)$ at the least favorable state of nature ϑ. This least favorable case

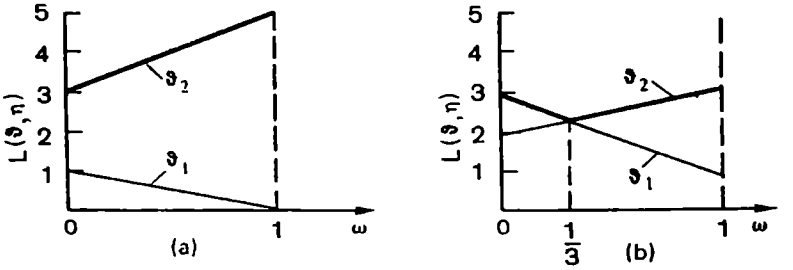


Fig. 10.3. Finding an optimal strategy by the minimax rule

will be that of $\vartheta \in \Theta$ at which $L(\vartheta, \eta)$ is a maximum. This quantity should be minimized by the statistician, who has to choose a minimax strategy $\eta^*(a)$ which meets the condition

$$L(\vartheta, \eta^*) = \min_{\eta} \max_{\vartheta} L(\vartheta, \eta) \qquad (10.10)$$

Example 10.4. Find a minimax strategy of the statistician in the processing line problem (Fig. 10.1). Since the solution must lie in the class of admissible strategies, then it will be sufficient to limit our evaluation to the strategies defined by the relations (10.8) and (10.9) and corresponding to the segments C_1C_2 and C_2C_3 of the convex hull S^* in Fig. 10.1. By these relations and line segments we construct strategy lines as the graphs of $L(\vartheta, \eta)$ versus w for $\vartheta = \vartheta_1$ and $\vartheta = \vartheta_2$ (Fig. 10.3). The values of $\max_{\vartheta} L(\vartheta, \eta)$ are emphasized by bold line. The minimum of this quantity is seen to be attained at $w = 0$ for C_1C_2 and is equal to 3, and on the C_2C_3 segment it is defined by the intersection of two lines $3 - 2w = 2 + w$, i.e. at

$w = 1/3$ and is equal to $7/3 < 3$. Thus, the minimax rule yields a point on C_2C_3 that corresponds to $w = 1/3$, i.e. defines an optimal mixed strategy $\eta^* = (0, 1/3, 2/3)$ for which the statistician will lose not more than 7/3 at any strategy of nature.

Sometimes it is rational to choose a strategy that is governed not by overall losses $L(\vartheta, a)$ but by the *additional* losses $L'(\vartheta, a)$ defined as

$$L'(\vartheta, a) = L(\vartheta, a) - \min_a L(\vartheta, a) \tag{10.11}$$

For every state of nature, the quantity $\min_a L(\vartheta, a)$ represents minimal loss to the statistician for his best action, i.e. a guaranteed loss. One may think that this loss should be compensated for in some way or another (say, by the respective rise of the prices) and consequently, need not necessarily be taken into account in strategy selection. In this case, a strategy may be chosen by applying the minimax rule to the additional losses.

Example 10.5. Table 10.2 summarizes the additional losses $L'(\vartheta, a)$ calculated by (10.11) in the processing line problem. If we apply the minimax rule to the entries of this table, we obtain, as an optimal strategy, a pure strategy a_1 with a loss of 1.

Table 10.2. $L'(\vartheta, a)$

	a_1	a_2	a_3
ϑ_1	0	1	3
ϑ_2	3	1	0

Whereas in strategic games the use of the minimax principle is fairly justified, from a statistical viewpoint treating nature as an intelligent opponent implies that the statistician is an extremely cautious player that is afraid most of all of suffering a casual large loss. The minimax approach is disadvantageous also because it fails to take into account prior information about states of nature and thus limits the gain that this information may help to win. Therefore, the minimax rules may be recommended where prior information about states of nature is absent or there are grounds to doubt that this information is true.

Another principle that takes into account the prior distribution of probabilities $\xi(\vartheta)$ is the Bayes principle. According to this approach, a mixed strategy $\eta(a)$ of the statistician is assessed by averaging the loss $L(\vartheta, \eta)$ over all possible states of nature $\vartheta \in \Theta$ observing the prior probability distribution $\xi(\vartheta)$, i.e. by obtaining the quantity

$$L(\xi, \eta) = \sum_{\vartheta} L(\vartheta, \eta)\, \xi(\vartheta) \tag{10.12}$$

A best strategy $\eta(a)$ will be that giving a minimum of $L(\xi, \eta)$. It is called the *Bayes strategy*.

The Bayesian approach may be applied naturally to both overall and additional losses. However, in examples that follow we shall limit ourselves to applying the Bayesian principle to overall loss only.

Example 10.6. In the processing line problem, for $\xi(\vartheta_1) = 0.6$ and $\xi(\vartheta_2) = 0.4$ and for admissible strategies defined by the segment C_1C_2 we have

$$L(\xi, \eta) = (1 - w)\cdot 0.6 + (3 + 2w)\cdot 0.4 = 1.8 + 2.0w$$

so that $\min_w L(\xi, \eta) = 1.8$ which at $w = 0$ corresponds to the mixed strategy $\eta = (0, 1, 0)$; while for admissible strategies defined by the segment C_2C_3 we have $L(\xi, \eta) = 2.6 - 0.8w$, so that $\min_w L(\xi, \eta) = 1.8$ which at $w = 1$ corresponds to the same mixed strategy $\eta = (0, 1, 0)$. Therefore, the Bayes strategy is the pure strategy a_2.

This result is not a chance one. We shall demonstrate that the Bayesian principle always yields one of pure strategies of the statistician as the best one. By the same token, the Bayesian action a^* will be given by the condition (10.3).

10.2.4 Geometric Interpretation of Bayesian Strategies

Consider a statistical S-game for two states of nature, ϑ_1 and ϑ_2, specified by the convex hull S^* on the (x, y) plane (Fig. 10.4), where

$$x = L(\vartheta_1, \eta) \quad \text{and} \quad y = L(\vartheta_2, \eta) \tag{10.13}$$

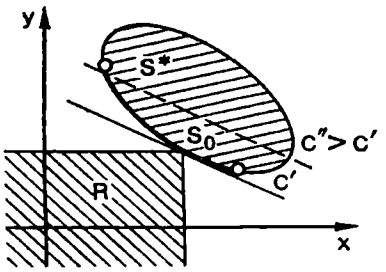


Fig. 10.4. Constructing a supporting line to a convex set

As indicated, the admissible strategies lie on the bottom left boundary of S^*. Consider one of such strategies, S_0, and construct a set R consisting of all points lying below and on the left of S_0. Clearly, S^* and R are convex sets and no one point of S^* belongs to R, including S_0.

Let us construct a line separating sets S^* and R. This line is to pass through S_0, i.e. it is a supporting line for S^* at S_0 and a supporting line for R. Consequently, this line must have a negative slope or be a vertical (horizontal) and its equation can be written as

$$y = -kx + c, \quad k \geqslant 0 \tag{10.14}$$

Dividing both sides of this equation by $k + 1$ reduces it to the form

$$ax + by = c' \tag{10.15}$$

where

$$a = k/(k + 1), \quad b = 1/(k + 1), \quad c' = c/(k + 1) \tag{10.16}$$

Recognizing that $a \geqslant 0$, $b \geqslant 0$, and $a + b = 1$, we may let $a = w$, $b = 1 - w$ and treat w and $1 - w$ as prior probabilities of states of nature ϑ_1 and ϑ_2, i.e.

$$w = \xi(\vartheta_1) \quad \text{and} \quad 1 - w = \xi(\vartheta_2) \tag{10.17}$$

The equation of the supporting line in this case can be written as

$$wx + (1 - w)\, y = c' \tag{10.18}$$

or

$$L(\vartheta_1, \eta)\, \xi(\vartheta_1) + L(\vartheta_2, \eta)\, \xi(\vartheta_2) = c' \tag{10.19}$$

Evidently, c' defines the expected losses $L(\xi, \eta)$ at the prior probabilities of states of nature equal to $\xi(\vartheta_1) = w$ and $\xi(\vartheta_2) = 1 - w$. It is not hard to note also that for point S_0, the value c' is minimal of all possible values of $L(\xi, \eta)$ since an increase of c' to $c'' > c'$ implies that the line will pass above and through points of S^* that are inadmissible, while a decrease of c' is impossible because the line will drop and will not have common points with S^* any longer. Therefore c' defines a strategy η providing a minimal expected loss $L(\xi, \eta)$ for a given distribution $\xi(\vartheta) = (w, 1 - w)$, i.e. for a given prior probability distribution defines a Bayesian strategy.

Because S_0 is an arbitrary point on the boundary of admissible strategies, then for any point of this boundary there exist such w and $1 - w$ for which this point will define a Bayesian strategy. We may conclude therefore that every admissible strategy is Bayesian for some prior probabilities w and $1 - w$.

Suppose now that prior probabilities of states of nature w and $1 - w$ are specified and we seek a point on the convex hull S^*, which would define a Bayesian strategy for this case. We construct in the (x, y) plane the line

$$wx + (1 - w)\, y = c \tag{10.20}$$

For an arbitrary c this line will be parallel to the supporting line at a point corresponding to a Bayes strategy. For convenience of construction we put $c = w(1 - w)$ and rewrite (10.20) as

$$x/(1 - w) + y/w = 1 \tag{10.21}$$

The line corresponding to this equation is shown in Fig. 10.5. Now, it is an easy matter to construct the supporting line that corresponds

to the given values of w and $1 - w$ and which defines on the convex hull S^* a point corresponding to a Bayesian strategy.

In this case the convex-hull boundary is a polygon whose vertices correspond to pure strategies of the statistician, therefore the supporting line has to pass through at least one vertex of the polygon. Consequently, given prior probabilities w and $1 - w$, there exists at least one Bayesian strategy being a pure strategy. This conclusion sizeably simplifies the solution of statistical games as the evaluation may be confined to a finite number of pure admissible strategies from infinite set of mixed strategies. This suits also many statisticians critical about the use of mixed strategies because these involve a randomization scheme irrelevant to the problem.

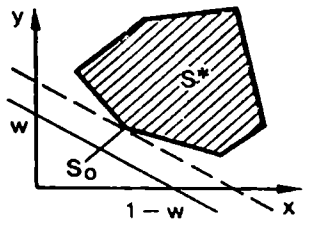


Fig. 10.5. Geometric evaluation of a Bayesian strategy

All these conclusions are valid also for the case where there are more than two states of nature. The geometrical consideration of such problems on the plane is, however, no longer possible.

10.3 Statistical Games with Sample Information

10.3.1 Problem Formulation

As has been already noted, in statistical games there exists a possibility for the statistician to improve and refine his knowledge on the state of nature by experimentation. If an unlimited experiment could be carried out, the statistician might obtain a perfect information about the state of nature and act under perfect certainty. However, any experiment is time and labor consuming, not to say of its cost which may be higher than the gain that could be obtained with the results of this experiment.

A sample experiment appreciably widens the class of statistician's strategies. First of all, the statistician is to decide whether or not an experiment is worth conducting. Then he is to decide what sort of experiment is necessary, how many trials should a sample involve, and what actions should be taken for some or other outcomes of the experiment.

We try to answer some of these questions in the ensuing sections. Until then we confine our attention to statistical games with a single experiment.

We assume that the decision about an experiment has been already made, and the single experiment we have in view is that whose size and order are defined in advance. For example, if we seek to check whether or not a coin is symmetric, we can carry out a single expe-

riment involving n tosses of the coin. The outcome of the experiment for $n = 5$ may be a sequence HTHHT. All in all there are 2^n such sequences, i.e. in this case the sample space of the experiments consists of 2^n elements. Likewise in testing a gun, a single experiment is the one including n shots. To estimate the effect of a drug on a rat, an experiment can be carried out in which daily response is measured in n test creatures for several months. Although in these examples a sample consists of a number of observations, we call it a single sample experiment as the nature of each trial is defined in advance.

10.3.2 Sample Space

Denote by Z the space of outcomes of the experiment with elements, i.e. individual outcomes, denoted $z_1, \ldots, z_v$.

Individual outcomes of the experiment $z \in Z$ find themselves connected with a state of nature $\vartheta \in \Theta$. This connection manifests itself in that for each state of nature $\vartheta \in \Theta$ there exists a probability $p(z \mid \vartheta)$ of obtaining a particular outcome $z \in Z$. The quantities $p(z \mid \vartheta)$ are the conditional probability distribution on space Z for a given ϑ with the common conditions

$$p(z \mid \vartheta) \geqslant 0, \quad \sum_z p(z \mid \vartheta) = 1 \tag{10.22}$$

The collection of three elements, a space of outcomes Z, space of states of nature Θ, and probability distribution $p(z \mid \vartheta)$ on Z at a given $\vartheta \in \Theta$ is called a *sample space*

$$\mathfrak{Z} = (Z, \Theta, p) \tag{10.23}$$

A sample space is convenient to specify in the form of a table designating the probability distribution $p(z \mid \vartheta)$ on the direct product $\Theta \times Z$.

Table 10.3. $p(z \mid \vartheta)$

	z_1	z_2	z_3	z_4
	Replacement problem			
ϑ_1	1/2	1/2	0	0
ϑ_2	0	1/2	1/2	0
ϑ_3	0	0	1/3	2/3
	Processing problem			
ϑ_1	0.60	0.25	0.15	
ϑ_2	0.20	0.30	0.50	

Example 10.7. In the tool replacement problem, an experiment may be a pre-operational test carried out by the works' personnel. Inadequate qualification of the personnel and lack of test equipment may result in the test data that reflect the state of the equipment under test only approximately. Suppose the experiment has four outcomes: z_1, the equipment is sound; z_2, maintenance is required; z_3, worn parts need be replaced; and z_4, the equipment is unsuitable for further use. The probabilities of these outcomes at various states of nature are summarized in Table 10.3 representing the sample space for this problem.

Example 10.8. In the processing line problem, an experiment may be a rough preliminary analysis of impurity content (an accurate laboratory analysis is impossible as it is time consuming and would cause equipment downtime). Results of this experiment may be as follows: z_1, no impurity found; z_2, impurities in low content; z_3, material is high in impurities. The respective sample space is displayed in Table 10.3.

10.3.3 Decision Rule

In a problem without sample information the statistician had to choose a decision from decision space A using only a prior information $\xi(\vartheta)$ on states of nature. In a problem with a sample test, the statistician makes a decision depending on the outcome of the experiment $z \in Z$. To formalize this problem, one may analyze all possible outcomes of the experiment in advance and set up a rule d that defines which decision should be made for each outcome $z \in Z$. This rule will be a mapping of the set of outcomes Z onto the decision space A

$$d: Z \to A \tag{10.24}$$

which may be written as

$$d(z) = a \tag{10.25}$$

The rule $d(z)$ suggesting a decision $a \in A$ that the statistician should make at any outcome of the experiment $z \in Z$ is called the *decision rule*. An example elucidating the concept of the decision rule follows.

Suppose that the decision space consists of three elements $A = \{a_1, a_2, a_3\}$ and the sample space consists of five observations $Z = \{z_1, z_2, z_3, z_4, z_5\}$. The decision rule $d(z_i) = a_i$ may be given as a set of pairs of indices (ij) implying a decision a_j for an observation z_i. The decision rule will be, for example, the set $\{(1, 1), (2, 1), (3, 2), (4, 2), (5, 3)\}$ implying that should the observations be z_1 and z_2, z_3 and z_4, and z_5, then the respective actions a_1, a_2 and a_3 are taken. Of course, this decision rule is not the only possible. Some other decision rule may be considered, say, $\{(1, 1), (2, 2), (3, 2), (4, 3), (5, 3)\}$ or $\{(1, 3), (2, 1), (3, 2), (4, 3), (5, 2)\}$ and the like. It would be convenient therefore to introduce a space D containing a complete list of all possible decision rules and called *space of decision rules.*

The decision rule considered above separates set Z into disjoint subsets S_{ai}, namely,

$$S_{a1} = \{z_1, z_2\}, \quad S_{a2} = \{z_3, z_4\}, \quad S_{a3} = \{z_5\} \tag{10.26}$$

In set-theoretic terms we may say that any decision rule $d \in D$ may be treated as a partition of sample space Z into disjoint subsets

S_a such that

$$S_a = \{z \mid d(z) = a\}, \; a \in A \tag{10.27}$$

Specifically, if decision space A consists only of elements a_1 and a_2, then the decision rule $d(z)$ partitions Z into a space S, called a *critical region*, and its complement $C(S) = \overline{S}$ defined by the condition

$$\begin{aligned} d(z) &= a_1 \quad \text{if} \quad z \in S \\ &= a_2 \quad \text{if} \quad z \in S \end{aligned} \tag{10.28}$$

The concept of decision rule allows a more explicit formulation of statistician's goal. To be more precise, he seeks to choose a decision rule $d(z) \in D$ which allows him to make most profitable decisions. To this end, however, he has to assess various decision rules and this can be done with the aid of a risk function.

10.3.4 Risk Function

If the statistician has chosen a decision rule $d(z)$, he thereby has specified for each outcome $z \in Z$ of the experiment a decision $a = d(z)$. For a given $\vartheta \in \Theta$ this decision incurs the loss

$$L(\vartheta, a) = L[\vartheta, d(z)] = L_z(\vartheta, d) \tag{10.29}$$

However, at a given ϑ the outcome z of the experiment will be a random variable defined by the probability distribution $p(z \mid \vartheta)$ on Z. Consequently, this probability will dictate the occurrence of losses $L_z(\vartheta, d)$ for this z, i.e. the loss is also a random function.

Because estimation of a decision rule $d(z)$ for a given state of nature ϑ has to take into account all possible outcomes of the experiment, we should speak of average losses over the entire space of outcomes Z. These expected losses are called the *risk function*

$$\rho(\vartheta, d) = E_z[L_z(\vartheta, d)] = \sum_z L_z(\vartheta, d)\, p(z \mid \vartheta) \tag{10.30}$$

Each state of nature $\vartheta \in \Theta$ and each decision rule $d \in D$ will be associated with its own value of expected loss, i.e. the risk function $\rho(\vartheta, d)$, which therefore is defined on the direct product of sets $\Theta \times D$ similar to how in the game without sample information the loss function $L(\vartheta, a)$ was defined on the direct product $\Theta \times A$. It follows then that the space of decision rules D and the risk function $\rho(\vartheta, d)$ play the same role in games with sample information as the decision space A and the loss function in games without experiments. This parallelism suggests similar methods of solution for these two cases.

In games with sample information the statistician may employ mixed strategies as well. To this end, he should have a random me-

chanism that specifies a probability distribution $\eta(d)$ on D. The respective risk function $\rho(\vartheta, \eta)$ for a mixed strategy $\eta(d)$ results then by averaging $\rho(\vartheta, d)$ over all pure strategies involved in this strategy. Thus,

$$\rho(\vartheta, \eta) = E_d[\rho(\vartheta, d)] = \sum_d \rho(\vartheta, d)\,\eta(d) \qquad (10.31)$$

or in view of (10.30),

$$\rho(\vartheta, \eta) = \sum_{z,\,d} L_z(\vartheta, d)\, p(z \mid \vartheta)\,\eta(d) \qquad (10.32)$$

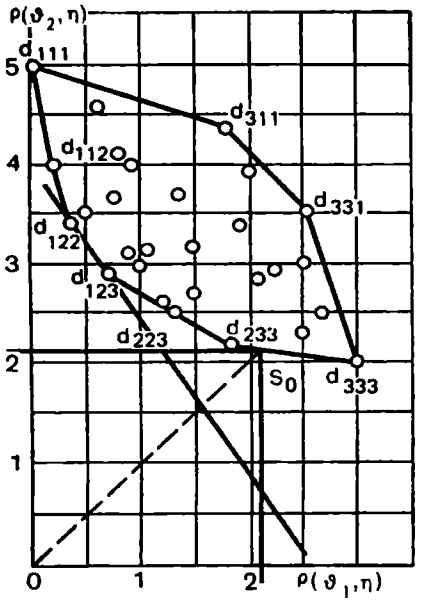


Fig. 10.6. Risk function in the processing line problem

It is quite natural that looking for a best strategy in a game with sample information the statistician should operate only with admissible strategies which are evaluated in the same manner as in games without sample information.

Example 10.9. Determine the risk function in the processing line problem. For convenience, the losses to the statistician $L(\vartheta, a)$ and the probabilities $p(z \mid \vartheta)$ of outcomes of the experiment, given in Tables 10.1 and 10.3, are summarized in Table 10.4.

Table 10.4

	$L(\vartheta, a)$			$p(z \mid \vartheta)$		
	a_1	a_2	a_3	z_1	z_2	z_3
ϑ_1	0	1	3	0.60	0.25	0.15
ϑ_2	5	3	2	0.20	0.30	0.50

Because the space of outcomes consists of three elements, then the decision rule is

$$d(z) = (a_i, a_j, a_k) = d_{ijk} \qquad (10.33)$$

where a_i, a_j, and a_k are decisions made by the statistician for the outcomes z_1, z_2, and z_3, respectively. For example, the decision rule d_{122} implies that for the outcomes z_1, z_2, and z_3 the statistician makes the decisions a_1, a_2, a_3.

The values of the risk function calculated by (10.30) for each decision rule are graphed in Fig. 10.6. An example of the compu-

tations will be given for $\rho(\vartheta, d_{122})$. By virtue of (10.30) we have

$$\begin{aligned}\rho(\vartheta, d_{122}) &= L_{z1}(\vartheta, d_{122})\, p(z_1 \mid \vartheta) + L_{z2}(\vartheta, d_{122}) \\ &\quad \times p(z_2 \mid \vartheta) + L_{z3}(\vartheta, d_{122})\, p(z_3 \mid \vartheta) \\ &= L(\vartheta, a_1)\, p(z_1 \mid \vartheta) + L(\vartheta, a_2)\, p(z_2 \mid \vartheta) \\ &\quad + L(\vartheta, a_2)\, p(z_3 \mid \vartheta)\end{aligned}$$

Letting $\vartheta = \vartheta_1$ and $\vartheta = \vartheta_2$, we get $\rho(\vartheta_1, d_{122}) = 0.4$ and $\rho(\vartheta_2, d_{122}) = 3.4$.

10.3.5 Evaluation of Strategies

Because the risk function reduces a game with sample information to the form similar to the case without experimentation, all solution techniques for the latter case are valid for this problem, the difference being that the statistician now has to minimize the average risk rather than the average losses.

The minimax principle here is to choose a strategy $\eta(d)$ such that would minimize the average risk $\rho(\vartheta, \eta)$ at the least favorable state of nature for the statistician. To put this another way, a minimax strategy η^* is chosen subject to the condition

$$\rho(\vartheta, \eta^*) = \min_{\eta} \max_{\vartheta} \rho(\vartheta, \eta) \tag{10.34}$$

If in evaluating the average risk $\rho(\vartheta, \eta)$ we take the additiona losses $L'(\vartheta, a)$ rather than overall losses, then (10.34) will define the minimax rule for additional losses.

In order to apply the Bayesian principle we introduce the notion of *expected risk* understood as an average risk taking into account all possible states of nature $\vartheta \in \Theta$ and the prior probability distribution on space Θ. Since in the Bayes approach the statistician may confine himself to using only pure strategies, the expected risk is

$$\rho(\xi, d) = \sum_{\vartheta} \rho(\vartheta, d)\, \xi(\vartheta) \tag{10.35}$$

The Bayes principle requires the use of a decision rule d^* such that minimizes the expected risk, called *Bayes*, namely,

$$\rho^*(\xi) = \rho(\xi, d^*) = \min_{d} \rho(\xi, d) \tag{10.36}$$

Example 10.10. Determine the minimax and Bayes strategies for the processing line problem with sample information. Figure 10.6 shows the S-game representation for this problem. As can be seen the minimax strategy (called the *minimax risk point* in the S-game setting) is given by point S_0 with the coordinates (30/14, 30/14) and corresponds to the pure strategies d_{333} and d_{333} randomized with the probabilities 10/14 and 4/14. The Bayes strategy will be d_{133}.

10.4 Use of Posterior Probabilities

10.4.1 Number of Strategies in Sample Information Games

The examples discussed in the preceding sections indicate that an experiment considerably increases the number of pure strategies of the statistician in statistical games. For example, in the processing line problem, the experiment raised the number of pure strategies from 3 to 27. If we failed to represent this game as a two-dimensional S-game, then its analysis would require a serious mathematical development.

Denote by N the total number of pure strategies of the statistician in a game with a sample test. This number is not hard to compute. Suppose that the game without sample information offers the statistician l decisions $a_1, \ldots, a_l$, and the number of possible outcomes of the experiment is v.

At $v = 1$ we obtain a game with the outcome of the experiment known in advance, which is equivalent to the game without this experiment when $N = l$.

At $v = 2$, the strategy of the statistician may be written as $d = (a_{i1}, a_{i2})$, where the decisions are any $a \in A$. Consequently, there may be l various a_{i1} to each of which there correspond l various a_{i2} so that $N = l^2$.

At $v = 3$, the statistician's strategy is $d = (a_{i1}, a_{i2}, a_{i3})$. The set of decisions a_{i1} and a_{i2} may assume l^2 values, each of which may be associated with l values of a_{i3}, so that $N = l^3$.

Continuing this argument leads to a conclusion that at arbitrary l and v the total number of pure strategies will be $N = l^v$. When l and v are large, the number of pure strategies of the statistician may be so large that their analysis may be really cumbersome.

Appreciable simplification of such situations can be achieved if in the evaluation of Bayes strategies the statistician employs the posterior distribution, computed by the results of the experiment, in place of the prior distribution. This approach reduces the number of pure strategies of the statistician to the level of the problem without sample test.

10.4.2 Posterior Probability Distribution

A prior probability distribution $\xi(\vartheta)$ derived on the basis of statistical data and past experience lends a useful information about the frequency of occurrence of a state of nature in general, i.e. irrespectively of those specific circumstances under which the statistician has to make a decision. The objective of an experiment performed by the statistician is to obtain an additional information about the true state of nature.

To get specific, assume that the sample space of the experiment is the set $Z = (z_1, \ldots, z_v)$. The outcome of a specific experiment will be random and therefore cannot be used for an accurate guess about the actual state of nature. This uncertainty, however, may be appreciably reduced by a correct design of the experiment.

The uncertainty about the state of nature can be altered because the experiment replaces the prior $\xi(\vartheta)$ with a new probability distribution $\xi(\vartheta \mid z)$, called the *posterior distribution*, on space Θ for a given outcome $z \in Z$.

A posterior probability distribution $\xi(\vartheta \mid z)$ can be found by the Bayes formula (5.26) which rewrites for our case as

$$\xi(\vartheta \mid z) = p(z \mid \vartheta)\, \xi(\vartheta)/p(z) \tag{10.37}$$

where $p(z \mid \vartheta)$ is the conditional probability of obtaining an outcome z at a given state of nature $\vartheta \in \Theta$, derived on the basis of statistical data, and

$$p(z) = \sum_{\vartheta} p(z \mid \vartheta)\, \xi(\vartheta) \tag{10.38}$$

is the unconditional probability of an observation z, which is obtained by the formula of total probability (5.25).

By virtue of (10.38) we rewrite the expression for posterior distribution in the form

$$\xi(\vartheta \mid z) = p(z \mid \vartheta)\, \xi(\vartheta) / \sum_{\vartheta} p(z \mid \vartheta)\, \xi(\vartheta) \tag{10.39}$$

Example 10.11. Determine the posterior distribution in the processing line problem. The calculation by (10.39) is summarized in Table 10.5. The calculation data are provided by the prior distribution $\xi(\vartheta)$ given by Table 10.1 and the conditional probability distribution $p(z \mid \vartheta)$ given in Table 10.3.

Table 10.5

ξ	$\xi(\vartheta)$	$p(z \mid \vartheta)$			$p(z \mid \vartheta)\, \xi(\vartheta)$			$\xi(\vartheta \mid z)$		
		z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3
ϑ_1	0.6	0.60	0.25	0.15	0.36	0.15	0.09	0.818	0.555	0.310
ϑ_2	0.4	0.20	0.30	0.50	0.08	0.12	0.20	0.182	0.445	0.690
$\sum_{\vartheta} p(z \mid v)\, \xi(v)$					0.44	0.27	0.29			

The determined posterior probabilities indicate that at the outcomes z_1 and z_3 the uncertainty with respect to states of nature sizeably reduces. However, at the outcome z_2 the uncertainty even increases to a level higher than the pre-experimental one.

10.4.3 The Maximum Likelihood Principle

The knowledge of posterior probabilities allows estimation of states of nature to be done with the aid of the maximum likelihood principle which takes the states recognized as most likely, on the basis of experimental data, as an estimate of the state of nature.

Example 10.12. We wish to guess the state of nature ϑ in the processing line problem for the outcome z_1. According to Table 10.5 the posterior probabilities of states of nature ϑ_1 and ϑ_2 for outcome z_1 are as follows:

$$\xi(\vartheta_1 \mid z_1) = 0.818 \quad \text{and} \quad \xi(\vartheta_2 \mid z_1) = 0.182$$

Since $\max[\xi(\vartheta_1 \mid z_1), \xi(\vartheta_2 \mid z_1)] = \xi(\vartheta_1 \mid z_1)$, then by the principle of maximum likelihood $\vartheta = \vartheta_1$.

The maximum likelihood approach is often used to choose a decision in the binary decision problem. The space of states of nature in this problem consists of two states $\Theta = \{\vartheta_1, \vartheta_2\}$ with the prior probability distribution $\xi(\vartheta) = (\zeta, 1 - \zeta)$. The space of decisions of the statistician also consists of two elements $A = \{a_1, a_2\}$, where a_1 represents a decision that nature is in state ϑ_1, and a_2 a decision of nature being in state ϑ_2. The probability distribution of outcomes for $\vartheta = \vartheta_1$ and ϑ_2 is also known. It is required to make a decision $a \in \{a_1, a_2\}$ basing on an outcome $z \in Z$ of the experiment.

Example 10.13. *The radar problem.* A mark on the radar screen can appear as a result of a reflection from an airplane or can be caused by noise. Consequently, there are two states of nature: ϑ_1 stands for the case when an object is present and ϑ_2 for the no object situation. By looking at the screen the operator may make two decisions: a_1, an object is there, and a_2, no object within radar's reach. Errors involved in this decision making are of two kinds:

$a_1 \mid \vartheta_2$, a type-one error or an error of the first kind, is sometimes referred to as the *false alarm* probability;

$a_2 \mid \vartheta_1$, a type-two error, or an error of the second kind, is sometimes referred to as the *miss probability*.

A decision in the binary decision problem is often made with the aid of the *likelihood ratio* defined by

$$\Lambda(z) = \xi(\vartheta_1 \mid z)/\xi(\vartheta_2 \mid z) \tag{10.40}$$

Given a number k chosen depending on the consequences to which an erroneous decision will lead, the likelihood-ratio decision rule is as follows:

take a_1 if $\Lambda(z) > k$;

take a_2 if $\Lambda(z) < k$;

take a_1 or a_2 if $\Lambda(z) = k$.

As for the significance of the consequences of an error, a false alarm for an antiairplane installation may have a far greater consequences than a miss probability. Therefore, action a_1 should be taken only if the decision maker is sure that an airplane is indeed present, i.e. if $\xi(\vartheta_1 \mid z) \gg \xi(\vartheta_2 \mid z)$. This implies that in this case k should be well above unity.

10.4.4 Bayes Rule Based on Posterior Probabilities

As indicated in the early sections of this chapter, the principal difficulty associated with the solution of posterior information problems consists in a sharp increase of the number of strategies with the number of possible outcomes of the experiment. However, in actual problems there is no need to consider all possible outcomes. If in the course of the experiment an outcome $z \in Z$ has been obtained, then the problem solution should be attempted for this particular outcome.

This can be done by calculating the posterior probability distribution $\xi(\vartheta \mid z)$ over state-of-nature space Θ for the given outcome z. This calculation will result in Θ, decision space A and the probability distribution $\xi(\vartheta \mid z)$ which has taken into account the result of the experiment. This problem differs from that without posterior information in the use of a posterior probability distribution $\xi(\vartheta \mid z)$ in place of a prior distribution $\xi(\vartheta)$. Consequently, the solution techniques for this problem are similar to that of prior information problem.

When posterior probabilities $\xi(\vartheta \mid z)$ are used, elements of decision space $A = \{a_1, \ldots, a_l\}$ are employed as pure strategies. Each action $a \in A$ is associated with the posterior loss, determined by analogy with (10.2),

$$L(\xi_z, a) = E_{\vartheta \mid z}[L(\vartheta, a)] = \sum_{\vartheta} L(\vartheta, a)\, \xi(\vartheta \mid z) \tag{10.41}$$

or in view of (10.39)

$$L(\xi_z \mid a) = \sum_{\vartheta} L(\vartheta, a)\, p(z \mid \vartheta)\, \xi(\vartheta) / \sum_{\vartheta} p(z \mid \vartheta)\, \xi(\vartheta) \tag{10.42}$$

A Bayes rule can be found by choosing, for each z, an action $a^* \in A$ which minimizes the posterior expected loss, defined by (10.42), or equivalently

$$R^*(\xi_z) = R(\xi_z, a^*) = \min_a (\xi_z, a) \tag{10.43}$$

10.4.5 Binary Decision Problem

We illustrate the principles outlined above by taking as an example a binary decision problem with a prior probability distribution $\xi(\vartheta) = (\zeta, 1 - \zeta)$. We consider losses under correct solution equal to zero. An error of the first kind $(a_2 \mid \vartheta_1)$ incurs a loss of w, while an error of the second kind $(a_1 \mid \vartheta_2)$ incurs a loss of unity. The loss matrix is as shown in Table 10.6.

Table 10.6

	a_1	a_2
ϑ_1	0	w
ϑ_2	1	0

Consider a decision rule $d(z)$ that according to (10.28) partitions the sample space Z into regions S and $\overline{S} = C(S)$ such that the decision a_1 is made if $z \in S$, and a_2 if $z \in C(S)$. Because S and $C(S)$ must be compact, the solution of this problem reduces to the evaluation of a boundary that separates Z into disjoint subsets S and $C(S)$. The elements $z \in Z$ composing

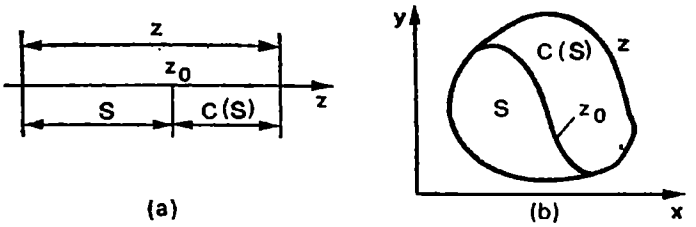


Fig. 10.7. Partition of the sample space in a binary decision problem

this boundary will be denoted z_0. Figure 10.7 illustrates the boundary z_0 for (a) one-dimensional and (b) two-dimensional cases.

To derive an equation describing z_0 we put down the expressions for average losses for given z and decisions a_1 and a_2. Using (10.42) and Table 10.6, we obtain

$$L(\xi_z, a_1) = p(z \mid \vartheta_2)(1 - \zeta) / \sum_{\vartheta} p(z \mid \vartheta)\, \xi(\vartheta) \tag{10.44}$$

$$L(\xi_z, a_2) = wp(z \mid \vartheta_1)\zeta / \sum_{\vartheta} p(z \mid \vartheta)\, \xi(\vartheta) \tag{10.45}$$

The boundary z_0 corresponds to the equality of the average losses at a_1 and a_2, which yields

$$(1 - \zeta)\, p(z_0 \mid \vartheta_2) = \zeta w p(z_0 \mid \vartheta_1) \tag{10.46}$$

or

$$p(z_0 \mid \vartheta_2)/p(z_0 \mid \vartheta_1) = \zeta w/(1-\zeta) = c(\zeta, w) \qquad (10.47)$$

As can be seen to every value of ζ there corresponds its own boundary z_0 and, consequently, the associated regions S_ζ and $C(S_\zeta)$. The quantity ζ will control also the probabilities of erroneous decisions $\alpha(\zeta)$ and $\beta(\zeta)$ that at $\vartheta = \vartheta_1$ point z will lie in $C(S_\zeta)$ and action a_2 will be taken, and that at $\vartheta = \vartheta_2$ point z will lie in S_ζ and action a_1 will be taken. These probabilities are defined by

$$\alpha(\zeta) = P(a_2 \mid \vartheta_1) = \sum_{z \in C(S_\zeta)} p(z \mid \vartheta_1) \qquad (10.48)$$

$$\beta(\zeta) = P(a_1 \mid \vartheta_2) = \sum_{z \in S_\zeta} p(z \mid \vartheta_2) \qquad (10.49)$$

To look at the character of these functions let us determine $\alpha(\zeta)$ and $\beta(\zeta)$ for the end points of $\zeta = 0$ and $\zeta = 1$.

The elements $z \in Z$ constituting the set S are defined by the condition $L(\xi_z, a_1) \leqslant L(\xi_z, a_2)$, or

$$p(z \mid \vartheta_2)/p(z \mid \vartheta_1) \leqslant \zeta w/(1-\zeta) \qquad (10.50)$$

which at $\zeta = 0$ becomes

$$p(z \mid \vartheta_2)/p(z \mid \vartheta_1) = 0$$

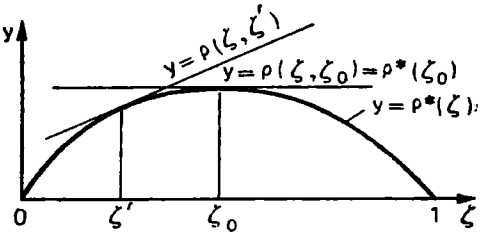


Fig. 10.8. Bayes risk in a binary decision problem

which is possible only if $S\zeta = \varnothing$ and $C(S\zeta) = z$. The formulas (10.48) and (10.49) yield $\alpha(0) = 1$ and $\beta(0) = 0$.

At $\zeta = 1$, (10.50) gives

$$p(z \mid \vartheta_2)/p(z \mid \vartheta_1) \leqslant \infty$$

which defines $S_\zeta = z$ and $C(S_\zeta) = \varnothing$, so that $\alpha(1) = 0$ and $\beta(1) = 1$.

Thus, as ζ varies from 0 to 1, $\alpha(\zeta)$ varies from 1 to 0, and $\beta(\zeta)$ from 0 to 1.

To determine the expected losses at all possible values of ζ, we evaluate the Bayes risk $\rho^*(\zeta)$ for the considered decision rule $d(z)$. Averaging the risk function of (10.30) over all states of nature, we find

$$\rho^*(\zeta) = \zeta w \alpha(\zeta) + (1-\zeta)\beta(\zeta) \qquad (10.51)$$

The plot of $\rho^*(\zeta)$ constructed by this formula and taking into account the behavior of $\alpha(\zeta)$ and $\beta(\zeta)$ is presented in Fig. 10.8. The plot is seen to vanish at $\zeta = 0$ and $\zeta = 1$ and peaks at some $\zeta = \zeta_0$. The value of ζ_0 defines the least favorable strategy of nature for the statistician at which his expected Bayesian loss is $\rho^*(\zeta_0)$.

A prior statistical information is often insufficient to formulate exactly the prior probability ζ. Therefore, it would be interesting to evaluate the Bayes risk for the case when the statistician uses some value of the prior probability ζ' and obtains values of S_ζ, $\alpha(\zeta')$ and $\beta(\zeta')$, whereas the true value of ζ will be different. In this case, the Bayes risk is defined by the expression

$$\rho(\zeta, \zeta') = \zeta w\alpha(\zeta') + (1 - \zeta)\beta(\zeta') \tag{10.52}$$

which is a linear function of ζ and describes a tangent to the curve $\rho^*(\zeta)$ at ζ'. The graph suggests that the loss incurred by the function $\rho(\zeta, \zeta')$ exceed that incurred by the function $\rho^*(\zeta)$ if $\zeta \neq \zeta'$ so that insufficient knowledge of the prior distribution leads to higher losses. Moreover, if ζ appreciably differs from ζ', then the loss defined by $\rho(\zeta, \zeta')$ may exceed $\rho^*(\zeta_0)$ corresponding to the maximum Bayes loss. However, if a least favorable probability distribution corresponding to ζ_0 is employed, then the function $y = \rho(\zeta, \zeta_0)$ will be constant and for any ζ will define losses equal to $\rho^*(\zeta_0)$. This case corresponds to employing a minimax strategy.

Thus, the Bayes approach is reasonable only in those situations where the prior distribution $\zeta(\vartheta)$ is known with sufficient accuracy. If the prior distribution is unknown or known without sufficient accuracy, then the minimax approach may prove more profitable.

10.5 Sequential Decision Analysis

10.5.1 Notation

In considering statistical games with a single experiment we noted that a single experiment need not necessarily consist of a single trial. It may involve a sequence of trials but the size and order of these trials should be defined in advance. For example, if we have agreed that a decision will be made as soon as N consecutive trials have been completed, then the entire sequence is treated as a single experiment whose outcome is a multivariate $z = (z_1, \ldots, z_N)$, where z_i for $i = 1, \ldots, N$, is the outcome of test i. Accordingly, a game with a single experiment is often referred to as a game with sample information or a game with a given sample size, the latter understood as the number of consecutively taken observations.

In games with a given sample size the statistician employs strategies that define his actions after the entire sequence of observations is taken. However, there is another way of sampling. Instead of completing all the N observations, he may decide, after each observation in the sequence is completed, whether sampling should be terminated and some decision should be chosen from A on the basis of the available information, or a next trial should be made. This

approach expands the class of possible strategies of the statistician, as the choice of decisions of A is augmented by the choice of a decision as to whether the experiment should be terminated or continued. This type of procedure may be called *games with sequential sampling*, but in the literature it is referred to as *sequential decision analysis*, or simply a *sequential procedure*. If the number of trials is limited by m observations after which a decision from A has to be made, then such a game is called an m-truncated procedure. We restrict our consideration in the following by such procedures referring the reader interested in more detailed discussions to Refs. [57, 62, and 63]. We shall refer to sequential results of an experiment as *observations*.

If the experiment had no cost, an expansion of the class of strategies by augmenting it with sequential sampling would not make sense, as the statistician without any loss may only gain from conducting all N observations. However, in many situations an experiment is a costly procedure not to say of time. Therefore a sensible statistician should compare at each stage of the experiment the cost of continuing the experiment with the expected gain due to the additional information. To lower the decision loss it will generally be necessary to run a larger experiment, whereby the experimental cost will be increased. In this section, we will be concerned with the balancing of these two costs.

To get specific, consider a description of a procedure with truncated sequential samples. Denote by Z_i the set of outcomes of test i. Then the sample space Z of the experiment after N tests may be represented as

$$Z = Z_1 \times \ldots \times Z_N \tag{10.53}$$

We assume that the sampling is stopped and a decision is made not after conducting all N observations $z_1, \ldots, z_n$, but only after j first observations $z_1, \ldots, z_j$. Therefore, the set Z can be partitioned into disjoint subsets $S_0, S_1, \ldots, S_N$ such that if $z \in S_j$, then a decision is made after first j observations.

The sequential sampling design $S = (S_0, S_1, \ldots, S_N)$ is called the *stopping rule*. It is frequently convenient to talk in terms of the stopping time, j, rather than the stopping rule. The stopping time is simply the final sample size, i.e. the sample size at which S says to stop sampling and make a decision.

The set Z may be partitioned into disjoint subsets S_j in several different ways. Each of these methods yields its own stopping rule. The set of all possible stopping rules (sampling designs) of a sequential procedure will be denoted $\mathfrak{S}$ and called in this text the *complete class of sequential samples.*

To make a decision, a sequential decision procedure $d(z)$ must be given which defines for every sequence of observations a decision

taken from A. Like in a game with a single experiment of a fixed sample size, the decision rule $d(z)$ is chosen from the set of decision functions D.

A strategy of the statistician in a sequential decision procedure consists, firstly, in choosing a stopping rule $S \in \mathfrak{S}$ indicating when the sampling must be stopped, and, secondly, in choosing a decision rule $d \in D$ indicating which decision must be made at this stopping. Thus, the pair (S, d) defines a strategy of the statistician. Observing that the pair (S, d) is an element of the direct product $\mathfrak{S} \times D$, we conclude that $\mathfrak{S} \times D$ is the space of pure strategies of the statistician.

The general number of strategies in sequential procedures is much greater than in fixed-sample-size problems so that a complete list of strategies for the former and choice of a best one are rather difficult to make. This procedure, however, may be simplified substantially by calculating, at each stage of sampling, a posterior probability distribution on the basis of all information collected by this instant. As the sampling proceeds, the uncertainty as to the true state of nature will be diminished and there appears a possibility to develop a criterion saying when to stop sampling and make a decision from the set A.

10.5.2 Bayesian Sequential Analysis

Here we shall consider only those statistical games whose observations $z_1, \ldots, z_N$ are independent random variables. Define $q(z_j \mid \vartheta)$ to be a probability distribution on Z_j at a given state of nature ϑ. The cost of an observation is assumed constant and equal to $c = 1$.

As a basis for description of a stopping rule we assume the probability distribution $\xi(\vartheta)$ on state-of-nature space Θ. Here, $\xi(\vartheta)$ will be understood as not only a prior distribution obtained before the experiment begins, but also as a posterior distribution obtained after several observations. Let $\xi_0(\vartheta)$ denote the prior distribution and $\xi_j(\vartheta)$ the posterior distribution after j observations.

The posterior distribution $\xi_j(\vartheta)$ will contain all information about the state of nature ϑ that was available before the experiment was started and obtained in the course of first j observations. Accordingly, $\xi_j(\vartheta)$ may be viewed as a prior distribution for the $(j+1)$th observation. Hence, the distribution $\xi_{j+1}(\vartheta)$ may be expressed through $\xi_j(\vartheta)$ with the aid of the formula for posterior probability distribution.

Consider the transition from the prior distribution $\xi_j(\vartheta)$ to the posterior distribution $\xi_{j+1}(\vartheta)$ as a transformation T applied to

$\xi_j(\vartheta)$. Then, according to (10.39), the posterior distribution is

$$T\xi_j(\vartheta) = \xi_{j+1}(\vartheta) = \xi_j(\vartheta)\, q(z_{j+1} \mid \vartheta) / \sum_{\vartheta} \xi_j(\vartheta)\, q(z_{j+1} \mid \vartheta) \quad (10.54)$$

We shall use a binary decision problem with $\Theta = \{\vartheta_1, \vartheta_2\}$ and $A = \{a_1, a_2\}$ to illustrate the idea underlying the derivation of a stopping rule on the basis of the posterior probability distribution. Denote by $(\xi, 1 - \xi)$ a posterior distribution on Θ after a few observations. The space E of mixed strategies of nature will be defined by the region of all possible values of ζ, i.e. by the interval $[0, 1]$ of the real axis.

It may so happen that $\zeta = 1$, then the decision to be made is a_1 only. If $\zeta = 0$, then we take action a_2. However, if $\zeta = 0.5$, then

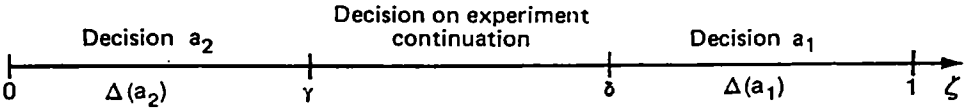


Fig. 10.9. Sequential decision procedure in a binary decision problem

it is impossible to prefer one decision to the other so that the experiment should be continued until the true state of nature will be refined.

These were the extreme cases of the distribution $\xi(\vartheta)$. In the general case, we may define two values $0 \leqslant \delta \leqslant 1$, and $0 \leqslant \gamma \leqslant 1$ $(\delta \geqslant \gamma)$ such that the statistician should

take a_1 if ζ is in the range $[\delta, 1]$;
take a_2 if ζ is in the range $[0, \gamma)$;
continue sampling if ζ lies in the interval (γ, δ).

Quite appropriately the intervals $\Delta(a_1) = [\delta, 1]$ and $\Delta(a_2) = [0, \gamma]$, graphed in Fig. 10.9, may be called *stopping* or *decision regions*. These regions may be defined also for the case when the number of states of nature is more than two. In this case, however, the space E of mixed strategies of nature will have a more complicated form. For example, for three states of nature, it has the form of an equilateral triangle of unit altitude (see Fig. 10.15).

In the general case, a stopping region $\Delta(a)$ in space E is a subset of E,

$$\Delta(a) \subset E \quad (10.55)$$

such that if after some observation $\xi(\vartheta) \in \Delta(a)$, then the sampling is stopped and the decision a is made.

Because every observation has some cost, then it will not be indifferent whether $\xi(\vartheta)$ appears in $\Delta(a)$ prior to the sampling or after a few observations. This implies that every new observation changes

both $\xi(\vartheta)$ and $\Delta(a)$. At each stage of sampling of interest are not those observations that have been already taken, but those that remain to be taken to the end of sampling. Therefore, we shall denote decision regions at stopping time j of N observations as $\Delta_{N-j}(a)$.

If the decision space consists of m elements $a_1, \ldots, a_m$, then at each state of sampling m regions $\Delta(a)$ must be determined that correspond to each $a \in A$. These regions must be convex and disjoint [57].

10.5.3 Sequential Decision Procedure

Suppose that m regions $\Delta(a)$ have been evaluated in space E at each stage of sampling. Then the sequential decision procedure will be as follows.

If the prior probability distribution $\xi_0(\vartheta) \in \Delta_N(a_i)$ at some i, then we make the decision a_i without sampling. If $\xi_0(\vartheta) \notin \Delta_N(a_i)$ at no one i, then the first observation is taken and the posterior distribution $\xi_1(\vartheta)$ is calculated. If $\xi_1(\vartheta) \in \Delta_{N-1}(a_1)$ at some i, then the sampling is stopped and the decision a_i is made. If $\xi_1(\vartheta) \notin \Delta_{N-1}(a_i)$ at no one i, then the next observation is taken.

In the general case, if the sampling was not stopped at the first $j-1$ observations, then an additional observation is taken and $\xi_j(\vartheta)$ is evaluated. If $\xi_j(\vartheta) \in \Delta_{N-i}(a)$ at some i, then the sampling is stopped and the decision a_i is taken. If choice has not been made after $N-1$ observations, then the final choice will be made after the Nth observation. To meet this condition, the regions $\Delta_0(a_i)$ should be selected such that

$$\bigcup_i \Delta_0(a_i) = E \tag{10.56}$$

As can be seen the Bayesian sequential problem reduces to one of determining regions $\Delta_{N-1}(a)$ in space E for all $a \in A$ and all j from 0 to N.

10.5.4 Risk Function for an Optimal Sequential Stopping Rule

At each stage of a sequential sampling, say after j observations, the risk function gives the minimal expected loss to the statistician who takes the best possible action including one of continuing the sampling experiment. Because the decision making is based on the knowledge of the posterior distribution $\xi_j(\vartheta)$, then the risk function will depend on this probability distribution. We denote the risk function after j observations $\rho^*(\xi_j)$.

We shall search for an expression describing the risk function in steps beginning from the last stage of the experiment. Assume that

all N observations have been taken and a posterior distribution $\xi_N(\vartheta)$ is found. The average loss to the statistician deciding $a \in A$ will then be

$$L(\xi_N, a) = \sum_{\vartheta} L(\vartheta, a)\, \xi_N(\vartheta) \tag{10.57}$$

The minimal value of this loss corresponding to the Bayes decision a^* will be

$$R^*(\xi_N) = L(\xi_N, a^*) = \min_a (\xi_N, a) \tag{10.58}$$

Because we can make no better decision than a^*, the quantity $R^*(\xi_N)$ coincides, at this stage, with the risk function

$$\rho^*(\xi_N) = R^*(\xi_N) \tag{10.59}$$

Consider now an arbitrary stage of sampling and evaluate the least expected loss that the statistician will incur after he has taken j observations ($j = 0, 1, \ldots, N-1$), having the distribution $\xi_j(\vartheta)$. It should be evident that the statistician can proceed as follows.

(1) Stop sampling and make $a \in A$. Then the minimal expected ioss is

$$R^*(\xi_j) = \min_a L(\xi_j, a) = \min_a \sum_{\vartheta} L(\vartheta, a)\, \xi_j(\vartheta) \tag{10.60}$$

(2) Decide to take the $(j+1)$st observation. The associated loss will be combined by the cost of sampling, that has been taken to be unity, and the loss of deciding best after $j+1$ observations, i.e. $\rho^*(\xi_{j+1})$.

Unfortunately the statistician does not know the probability distribution $\xi_{j+1}(\vartheta)$. He knows only the distribution $\xi_j(\vartheta)$, and $\xi_{j+1}(\vartheta)$ can be evaluated by (10.54) as $T\xi_j(\vartheta)$. The value of $T\xi_j(\vartheta)$ can be evaluated for specific values $z_{j+1} \in Z_{j+1}$ and $\vartheta \in \Theta$ unknown to the statistician thus far, so that he may speak not of a specific value of the risk function $\rho^*(\xi_{j+1})$ but rather of the expected value $E[\rho^*(T\xi_j)]$, the averaging must be carried out over all possible values $z_{j+1} \in Z_{j+1}$ taken from the density $q(z_{j+1} \mid \vartheta)$ and over all $\vartheta \in \Theta$ taken from the density $\xi_j(\vartheta)$, viz.,

$$E[\rho^*(T\xi_j)] = \sum_{\vartheta,\, z_{j+1}} \rho^*[T\xi_j(\vartheta)]\, q(z_{j+1} \mid \vartheta)\, \xi_j(\vartheta) \tag{10.61}$$

Therefore, having taken the decision to continue sampling, the minimal loss to the statistician will be

$$1 + E[\rho^*(T\xi_j)] \tag{10.62}$$

Having taken j observations, the risk function will be defined by the minimal expected loss of the alternative to stop and to continue

sampling, namely,

$$\rho^* (\xi_j) = \min \{R^* (\xi_j), \quad 1 + E [\rho^* (T\xi_j)]\} \tag{10.63}$$

This expression can be written in a more compact form that does not involve the index of observation j. Consider a sampling stage when k stages are to go to the end of the experiment. Denote by $\xi (\vartheta)$ the posterior distribution at this state and by $\xi' (\vartheta) = T\xi (\vartheta)$ the posterior distribution after one more observation will be taken. Define $\rho_k^* (\xi)$ to be the risk function with k observations to go to the end of sampling. Then (10.59) and (10.63) rewrite as

$$\rho_0^* (\xi) = R^* (\xi) \tag{10.64}$$

$$\rho_k^* (\xi) = \min \{R^* (\xi), \ 1 + E [\rho_{k-1}^* (T\xi)]\} \tag{10.65}$$

Denote by V the sample space of the next observation, whose elements will be v. In this notation we get

$$R^* (\xi) = \min_a \sum_{\vartheta} L (\vartheta, \ a) \xi (\vartheta) \tag{10.66}$$

$$E [\rho_{k-1}^* (T\xi)] = \sum_{\vartheta, v} \rho_{k-1}^* [T\xi (\vartheta)] \, q (v \mid \vartheta) \, \xi (\vartheta) \tag{10.67}$$

$$T\xi (\vartheta) = \xi (\vartheta) \, q (v \mid \vartheta) / \sum_{\vartheta} \xi (\vartheta) \, q (v \mid \vartheta) \tag{10.68}$$

The formulas (10.64) and (10.65) may be used for the evaluation of stopping regions. However, comparatively simple results can be obtained for the binary decision problem only, therefore we restrict our consideration to this case.

10.5.5 Evaluation of Stopping Regions for a Binary Decision Problem with Limited Stopping Time

Consider a binary decision problem specified by its loss matrix

$$\mathbf{Q} = \begin{pmatrix} 0 & q_{12} \\ q_{21} & 0 \end{pmatrix} \tag{10.69}$$

and the probability distribution $\xi (\vartheta) = (\zeta, \ 1 - \zeta)$ on state-of-nature space. Denote by $\Delta_k (a_1)$ and $\Delta_k (a_2)$ the stopping regions corresponding to decisions a_1 and a_2 when there remain k observations to the end of the experiment. These regions are defined by values $\zeta = \delta_k$ and $\zeta = \gamma_k$ $(\gamma_k \leqslant \delta_k)$ such that

$$\Delta_k (a_1) = [\delta_k \leqslant \zeta \leqslant 1] \tag{10.70}$$

$$\Delta_k (a_2) = [0 \leqslant \zeta \leqslant \gamma_k]$$

The evaluation of stopping regions reduces to finding the values γ_k and δ_k for $k = 0, \ 1, \ \ldots, \ N$. This can be done by making use of the expressions derived for the risk function.

We begin with the evaluation of the boundary points of stopping regions for $k = 0$. Expressing $\xi(\vartheta)$ in terms of ζ in (10.64) and (10.66), we have

$$\rho_0^*(\zeta) = R^*(\zeta) = \min[(1 - \zeta) q_{21}, \zeta q_{12}] \tag{10.71}$$

By virtue of (9.56), $\Delta_0(a_1)$ and $\Delta_0(a_2)$ must be disjoint real-valued sets whose union gives the closed interval [0, 1]. Consequently, these regions must be separated from one another by a point $\zeta = \gamma_0 = \delta_0$ which may be found from the condition

$$(1 - \zeta) q_{21} = \zeta q_{12} \tag{10.72}$$

This expression represents the equality of the minimal expected losses of making decisions a_1 and a_2. From (10.72) it follows that

$$\gamma_0 = \delta_0 = q_{21}/(q_{12} + q_{21}) \tag{10.73}$$

For arbitrary $k \neq 0$ the risk function

$$\rho_k^*(\zeta) = \min\{R^*(\zeta), 1 + E[\rho_{k-1}^*(T\zeta)]\} \tag{10.74}$$

The boundary values ζ for the sets $\Delta_k(a_1)$ and $\Delta_k(a_2)$ can be obtained from the equality of the minimal expected losses of deciding a_1 or a_2 or deciding to continue sampling, namely,

$$R^*(\zeta) = 1 + E[\rho_{k-1}^*(T\zeta)] \tag{10.75}$$

If we observe that in (10.71) the quantity ζq_{21} is a monotonously increasing function of ζ vanishing at $\zeta = 0$ and $(1 - \zeta) q_{12}$ is a mo-

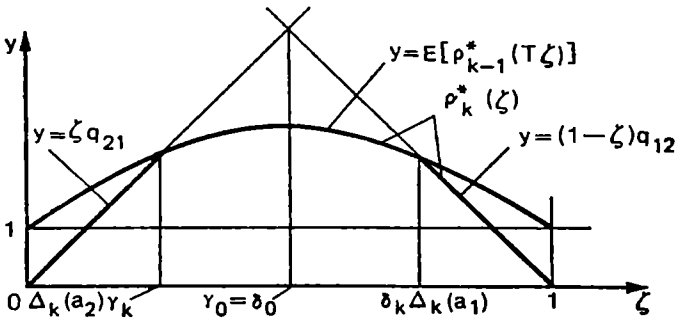


Fig. 10.10. Evaluation of stopping regions in a binary decision problem

notonously decreasing function of ζ vanishing at $\zeta = 1$, and replace at the boundary points ζ with γ_k for $\Delta_k(a_2)$ and with δ_k for $\Delta_k(a_1)$, the condition (10.75) may be written in the form

$$\gamma_k q_{21} = 1 + E[\rho_{k-1}^*(T\gamma_k)] \tag{10.76}$$

$$(1 - \delta_k) q_{12} = 1 + E[\rho_{k-1}^*(T\delta_k)] \tag{10.77}$$

These conditions are graphed in Fig. 10.10.

The expressions thus derived indicate that the evaluation of γ_k and δ_k and the risk function $\rho_k^*(\zeta)$ boil down to the calculation of

$E[\rho^*_{k-1}(T\zeta)]$. The formula (10.67) is a general expression for this quantity at an arbitrary Θ. For a binary decision problem this expression takes the form

$$E[\rho^*_{k-1}(T\zeta)] = \zeta \sum_v \rho^*_{k-1}(T\zeta)\, q(v \mid \vartheta_1) + (1-\zeta) \sum_v \rho^*_{k-1}[T(1-\zeta)]\, q(v \mid \vartheta_2) \tag{10.78}$$

where

$$T\zeta = \zeta q(v \mid \vartheta_1)/[\zeta q(v \mid \vartheta_1) + (1-\zeta)\, q(v \mid \vartheta_2)] \tag{10.79}$$

$$T(1-\zeta) = \frac{(1-\zeta)\, q(v \mid \vartheta_2)}{\zeta q(v \mid \vartheta_1) + (1-\zeta)\, q(v \mid \vartheta_2)} \tag{10.80}$$

At $k = 1$ the function $\rho^*_0(\zeta)$ defined by (10.71) is a very simple function of ζ, so that $E[\rho^*_0(T\zeta)]$ may be determined directly from (10.78). Given $E[\rho^*_0(T\zeta)]$, $\rho^*_1(\zeta)$ can be obtained by (10.74) and, consequently, $E[\rho^*_1(T\zeta)]$ can be obtained by (10.78), then ρ^*_2 by (10.74) and so on. Although this procedure involves long computations in evaluating $\rho^*_k(\zeta)$, they are not more complicated than taking of mathematical expectation.

10.6 Parameter Estimation and Filtering

10.6.1 Estimation Problems

Estimation problems are an extension of multiple decision problems discussed in preceding sections. One problem of practical value concerns estimation of vectors of states and system parameters [64-66]. It can be formulated as follows. Suppose that we wish to estimate the value of a scalar or vector parameter x that cannot be measured directly. Therefore, we have to measure another parameter z depending on x. The estimation problem seeks to answer the question: what can be said of x observing z?

Assume that x is an l-component vector and z is an m-vector. We shall call the value of x that we determine by this method the *estimate* $\hat{\mathbf{x}}$ of x. The rule which we use to find $\hat{\mathbf{x}}$ from z is the *estimator*. In other words, an estimator is the rule or procedure for processing a set of observations to generate an estimate. The estimator is the parallel of the decision rule, while the estimate is the parallel of the decision. Several methods for estimating x are possible. In order to judge which of the possible vectors $\hat{\mathbf{x}}$ estimates x best of all we need to develop a criterion J that would measure the quality of $\hat{\mathbf{x}}$.

Such a measure may be built around the estimation error $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$ as some nonnegative function $g(\mathbf{x})$ which by analogy with

(10.1) may be referred to as a loss or cost function. A natural choice for such a criterion is the expectation of the loss function $g(\tilde{\mathbf{x}})$ given the observation $\mathbf{z}$

$$J = E_{\mathbf{x}|\mathbf{z}}[g(\tilde{\mathbf{x}})] = \int_{\mathbf{x}} g(\tilde{\mathbf{x}})\, w(\mathbf{x} \mid \mathbf{z})\, d\mathbf{x} \tag{10.81}$$

where $w(\mathbf{x} \mid \mathbf{z})$ is the conditional probability density of $\mathbf{x}$ given the observation $\mathbf{z}$.

Following the Bayesian rule, we should take as an optimal estimate an $\hat{\mathbf{x}}$ such that minimizes J, i.e. satisfies the condition

$$\frac{\partial J}{\partial \hat{\mathbf{x}}} = \int_{\mathbf{x}} \frac{\partial g(\tilde{\mathbf{x}})}{\partial \hat{\mathbf{x}}}\, w(\mathbf{x} \mid \mathbf{z})\, d\mathbf{x} = 0 \tag{10.82}$$

It is quite natural that the optimal estimate depends on the chosen type of loss function. The more popular Bayes estimators are the *mean square cost function*

$$g(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}} = (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \hat{\mathbf{x}}) \tag{10.83}$$

and the *absolute value cost function*

$$g(\tilde{\mathbf{x}}) = \| \tilde{\mathbf{x}} \| = (\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}})^{1/2} \tag{10.84}$$

We restrict our consideration to the mean square error for which, in agreement with (4.115), we find

$$\frac{\partial g(\tilde{\mathbf{x}})}{\partial \hat{\mathbf{x}}} = -2(\mathbf{x} - \hat{\mathbf{x}})$$

so that by virtue of (10.82) the optimal estimate satisfies the relationship

$$\int_{\mathbf{x}} (\mathbf{x} - \hat{\mathbf{x}})\, w(\mathbf{x} \mid \mathbf{z})\, d\mathbf{x} = 0$$

or

$$\hat{\mathbf{x}} \int_{\mathbf{x}} w(\mathbf{x} \mid \mathbf{z})\, d\mathbf{x} = \int_{\mathbf{x}} \mathbf{x} w(\mathbf{x} \mid \mathbf{z})\, d\mathbf{x}$$

Recognizing that the integral in the left-hand side equals unity and the right-hand side is the conditional mean of $\mathbf{x}$, we obtain

$$\hat{\mathbf{x}} = E(\mathbf{x} \mid \mathbf{z}) \tag{10.85}$$

If both $\mathbf{x}$ and $\mathbf{z}$ are normally distributed with the means $E(\mathbf{x}) = \bar{\mathbf{x}}$, $E(\mathbf{z}) = \bar{\mathbf{z}}$ and dispersion (variance-covariance) matrices $\mathbf{D}_{xx} =$

$E[(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T]$, $\mathbf{D}_{zz} = E[(\mathbf{z}-\bar{\mathbf{z}})(\mathbf{z}-\bar{\mathbf{z}})^T]$, and $\mathbf{D}_{xz} = E[(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{z}-\bar{\mathbf{z}})^T]$, then in agreement with (10.85) the optimal estimate will be defined by (5.114) as

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{D}_{xz}\mathbf{D}_{zz}^{-1}(\mathbf{z}-\bar{\mathbf{z}}) \tag{10.86}$$

The relation (5.115) gives the dispersion matrix of estimate

$$Q = E[(\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})^T] = \mathbf{D}_{xx} - \mathbf{D}_{xz}\mathbf{D}_{zz}^{-1}\mathbf{D}_{zx} \tag{10.87}$$

Example 10.14. A scalar quantity x, that may with uniform probability be found in the interval [1, 11], is measured in the presence of normally distributed noise s with zero mean and $\sigma_s^2 = 1$. An observation $z = x + s$ gave 8. What is the estimate of x?

The expectation and variance for x will be (see Examples 5.9 and 5.11) $\bar{x} = (11+1)/2 = 6$ and $\sigma_x^2 = (11-1)^2/12 = 8.3$. In order to apply the above procedure we replace the uniform law for x with a normal distribution of the same mean and variance. Because x and z are scalars and $\bar{z} = \bar{x}$, the dispersion matrices of (10.86) and (10.87) become the variances and covariances

$$D_{xx} = E[(x-\bar{x})^2] = \sigma_x^2$$
$$D_{xz} = E[(x-\bar{x})(z-\bar{z})] = E[(x-\bar{x})(x-\bar{x}+s)] = \sigma_x^2$$
$$D_{zz} = E[(z-\bar{z})^2] = E[(x-\bar{x}+s)^2] = \sigma_x^2 + \sigma_s^2$$

By (10.86) the optimal estimate is

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \sigma_x^2(z-z)/(\sigma_x^2+\sigma_s^2) = 6 + 8.3(8-6)/9.3 = 7.78$$

By (10.87) the variance of this estimate is

$$Q = \sigma_x^2 - \sigma_x^4/(\sigma_x^2+\sigma_s^2) = \sigma_x^2\sigma_s^2/(\sigma_x^2+\sigma_s^2) = 8.3 \times 1 \div 9.3 = 0.89$$

It would be instructive to note that if we let $\sigma_x^2 = \infty$, which means that no prior information about x is available, then the optimal estimate is $\hat{x} = z = 8$ with variance $Q = \sigma_s^2 = 1$. This means that prior information would allow an estimate with a smaller variance, i.e. more accurate.

10.6.2 Least-Squares Method

The Bayes approach considered above requires a complete stochastic information for the parameter to be estimated, which is often hard to obtain. Here we consider the least-squares approach which uses no stochastic information but instead treats the parameter estimation task as a deterministic optimization problem.

We assume a linear observation model of the form

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{10.88}$$

where $\mathbf{n}$ is noise. The parameter $\mathbf{x}$ and noise $\mathbf{n}$ are assumed to be uncorrelated. The dispersion matrix of noise $\mathbf{D}_n = E[\mathbf{n}\,\mathbf{n}^T]$ is given by (5.163).

Let $\hat{\mathbf{x}}$ be some estimate of $\mathbf{x}$. The magnitude of the error $\boldsymbol{\varepsilon} = \mathbf{z} - \mathbf{H}\hat{\mathbf{x}}$ depends on the estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$. According to the least-squares method an optimum estimator $\hat{\mathbf{x}}$ is such that minimizes the squared error weighted by the weighting matrix $\mathbf{W} = \mathbf{D}_n^{-1}$. The problem is then to select an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ such that the quadratic performance index

$$J = \frac{1}{2}\,\boldsymbol{\varepsilon}^T\mathbf{D}_n^{-1}\boldsymbol{\varepsilon} = \frac{1}{2}\,(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}})\,\mathbf{W}\,(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}) \tag{10.89}$$

is minimized. The weighting matrix $\mathbf{W}$ is assumed to be positive definite and symmetric. This weighting matrix may be used to assign different costs to each of the errors $(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}})_i$. The estimator which minimizes $J(\hat{\mathbf{x}})$ is known as the *least-squares estimator*. This method is sometimes referred to as *weighted least squares* because of the use of the weighting matrix $\mathbf{W}$.

A necessary condition for the least-squares estimator is

$$\frac{\partial J}{\partial \hat{x}} = \mathbf{H}^T\mathbf{W}\,(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}) = 0$$

whence

$$\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{z} \tag{10.90}$$

If the disturbance is uncorrelated (white noise), then in agreement with (5.164) the dispersion matrix of noise in (10.89) is a unit matrix $\mathbf{I}$ and

$$\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{z} \tag{10.91}$$

Example 10.15. The readings of inaccurate devices in the electrical circuit shown in Fig. 10.11 are $A_1 = 1$ A, $A_2 = 2$ A, and $V = 20$ V. All resistors are $R = 5\ \Omega$. Assuming that the measurement errors have the same variance and are uncorrelated, it is required to find the estimates of currents i_1 and i_2.

The readings are related to i_1 and i_2 by the system of equations derived with Kirchhoff's laws

$$z_1 = A_1 = i_1 + \xi_1 = 1$$
$$z_2 = A_2 = i_1 + i_2 + \xi_2 = 2$$
$$z_3 = V/R = i_1 + 2i_2 + \xi_3 = 4$$

In vector form this system becomes $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{n}$ or

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} i_1 \\ i_2 \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}$$

where

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} i_1 \\ t_2 \end{pmatrix}, \quad \mathbf{n} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}$$

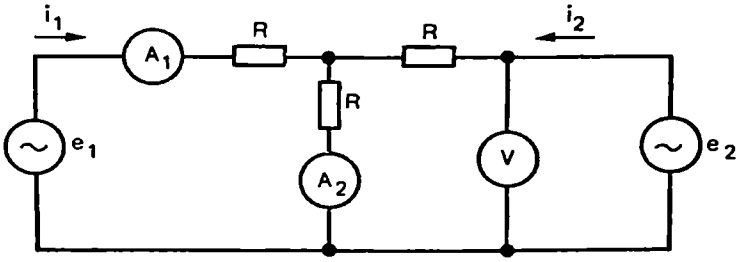


Fig. 10.11. Electrical circuit with inaccurate measuring devices

Because the values of noise are uncorrelated, the least-squares estimate will be found by (10.91) in which

$$\mathbf{H}^{\mathrm{T}}\mathbf{H} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}$$

$$(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} = \frac{1}{6} \begin{pmatrix} 5 & -3 \\ -3 & 3 \end{pmatrix}$$

$$\mathbf{H}^{\mathrm{T}}\mathbf{z} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 7 \\ 10 \end{pmatrix}$$

Equation (10.91) becomes

$$\hat{\mathbf{x}} = \frac{1}{6} \begin{pmatrix} 5 & -3 \\ -3 & 3 \end{pmatrix} \begin{pmatrix} 7 \\ 10 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 5 \\ 9 \end{pmatrix}$$

Thus, $\hat{i}_1 = 5/6 = 0.833$ A and $\hat{i}_2 = 9/6 = 1.5$ A.

10.6.3 Performance of the Least Squares Estimator

Analysis of accuracy of the least-squares estimator boils down to the evaluation of the expected value of the estimation error $E(\tilde{\mathbf{x}})$ and the variance matrix $Q = E[(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}}]$. From the

expression (10.90) for the least-squares estimator $\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1} \times \mathbf{H}^T\mathbf{W}\mathbf{z} = \mathbf{K}\mathbf{z}$ so that

$$\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}} = \mathbf{x} - \mathbf{K}\mathbf{z} = \mathbf{x} - \mathbf{K}(\mathbf{H}\mathbf{x} + \mathbf{n})$$

Since $\mathbf{K}\mathbf{H} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{H} = \mathbf{I}$, then $\tilde{\mathbf{x}} = -\mathbf{K}\mathbf{n}$. Hence we see that if $E(\mathbf{n}) = 0$, then $E(\tilde{\mathbf{x}}) = -\mathbf{K}E(\mathbf{n}) = 0$, which implies that the estimator $\hat{\mathbf{x}}_{LS}$ is unbiased, i.e. that $E(\hat{\mathbf{x}}) = E(\mathbf{x})$.

The variance (dispersion) matrix of $\hat{\mathbf{x}}$ is

$$Q = E(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T) = \mathbf{K}E(\mathbf{n}\mathbf{n}^T)\mathbf{K}^T = \mathbf{K}\mathbf{D}_n\mathbf{K}^T = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1} \times [\mathbf{H}^T\mathbf{W}\mathbf{W}^{-1}(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}]^T$$

Since $\mathbf{W}\mathbf{W}^{-1} = \mathbf{I}$ and both $(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}$ and $\mathbf{W}$ are symmetric matrices that do not change in transposition, we have

$$Q = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{H}^T\mathbf{W})^T(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{H}(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}$$

or

$$Q = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1} \tag{10.92}$$

If the values of noise are uncorrelated and have the same variance in each measurement, then $\mathbf{W}^{-1} = \mathbf{D}_n = E(\mathbf{n}\mathbf{n}^T) = \sigma^2\mathbf{I}$ and

$$Q = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1} \tag{10.93}$$

10.6.4 Nonlinear Estimation with Least Squares

Throughout this section it has been assumed that the observation is a linear function of the parameter vector (message). When it is a nonlinear function of the message, the problem of forming an optimal estimate becomes quite complex. To get specific, let $\mathbf{x} = (x_1, \ldots, x_l)^T$ be the vector of estimated parameters and $z_i = f_i(x_1, \ldots, x_l) + \xi_i = f_i(\mathbf{x}) + \xi_i$, for $i = 1, \ldots, n$, be a set of observations. If $f_i(x_1, \ldots, x_n)$ is a nonlinear function of the estimated parameters, then the functional of the least-squares method

$$J = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}[z_i - f_i(x_1, \ldots, x_n)]^2$$

becomes so involved that the system of equations $\partial J/\partial x_i = 0$, $i = 1, \ldots, n$, in most cases cannot be handled by analytical methods. However, if at least a rough approximation $\mathbf{x}_a$ to the required solution is obtained (rough estimate), then an iteration technique may be brought in to refine the solution.

Letting $\mathbf{x} = \mathbf{x}_a + \Delta\mathbf{x}$, where $\mathbf{x}_a$ is a rough estimate of $\mathbf{x}$ determined in some way, and expanding $f_i(\mathbf{x})$ in the Taylor series at the po-

int $\mathbf{x}_a$ we obtain by keeping the linear terms

$$f_i(\mathbf{x}) = f_i(\mathbf{x}_a + \Delta\mathbf{x}) = f_i(\mathbf{x}_a) + f_i'(\mathbf{x}_a)\,\Delta\mathbf{x}$$

where

$$f_i'(\mathbf{x}_a) = \frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}}\bigg|_{\mathbf{x}=\mathbf{x}_a}$$

We write the set of observations z_i for $i = 1, \ldots, n$ as

$$z_i = f_i(\mathbf{x}_a) + f_i'(\mathbf{x}_a)\,\Delta\mathbf{x} + \xi_i$$

Denoting by $\delta z_i = z_i - f_i(\mathbf{x}_a)$, we have

$$\delta z_i = f_i'(\mathbf{x}_a)\,\Delta\mathbf{x} + \xi_i, \text{ for } i = 1, \ldots, n$$

This system of equations is linear in the correction vector $\Delta\mathbf{x}$ and allows solution by least squares. In vector form this system of equations is

$$\Delta\mathbf{Z} = \mathbf{H}\Delta\mathbf{x} + \mathbf{n}$$

where $\Delta\mathbf{Z} = (\delta z_1, \ldots, \delta z_n)^{\mathrm{T}}$, $\mathbf{H} = (f_1(\mathbf{x}_a), \ldots, f_n'(\mathbf{x}_a))^{\mathrm{T}}$, and $\mathbf{n} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$.

The estimate of the correction vector is

$$\widehat{\Delta\mathbf{x}} = (\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{W}\Delta\mathbf{Z}$$

and the refined value of the estimate will be $\hat{\mathbf{x}} = \mathbf{x}_a + \widehat{\Delta\mathbf{x}}$. Now, $\hat{\mathbf{x}}$ may be taken as a new rough estimate, $\mathbf{x}_a'$, and, repeating the entire cycle, we get a better estimate, and so on. A criterion for convergence of the iteration process may be the condition $\widehat{\Delta\mathbf{x}} \approx 0$.

10.6.5 Filtering Problem Statement

In the remaining portion of this section we shall consider a new variety of estimation problems, known as *state estimation*. Whereas we estimated above a constant parameter $\mathbf{x}$ now we wish to estimate the time varying state of a dynamic system or process, i.e. evaluate current estimates of a time-varying scalar or vector signal $\mathbf{u}(t)$ corrupted by noise and therefore inaccessible for direct observation. The term *filtering* is used normally for estimations of a current value of $\hat{\mathbf{u}}(t)$, whereas the estimation is referred to as *prediction* or *extrapolation* for future values and as *smoothing* or *interpolation* for past values of the signal [66-68].

A specific form of the algorithms used to solve filtering problems depends on the stochastic properties of the signal and noise. Consider a system in which the signal is a slowly varying function of time and the interference is a random process of the type of white noise.

Current values of the signal can be estimated by Bayes' rule and by least squares. Here we shall focus on the latter technique. Assume

that the signal is a slowly varying scalar process $u(t)$ that may be represented on the finite observation interval as a Taylor series with a finite number of $(r+1)$ terms. Such a signal is represented in Fig. 10.12, and its expansion in the Taylor series at the current time instant t has the form

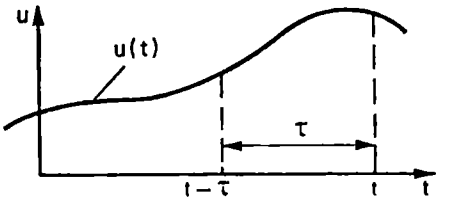


Fig. 10.12. Slowly varying signal

$$u(t-\tau) = u(t) - \tau\dot{u}(t) + \frac{\tau^2}{2}\ddot{u}(t) + \frac{(-\tau)^r}{r!}u^{(r)}(t) \qquad (10.94)$$

This relationship is a polynomial model of the signal, that at $r = 0, 1, 2$, etc. represents a constant signal, a ramp signal, and so

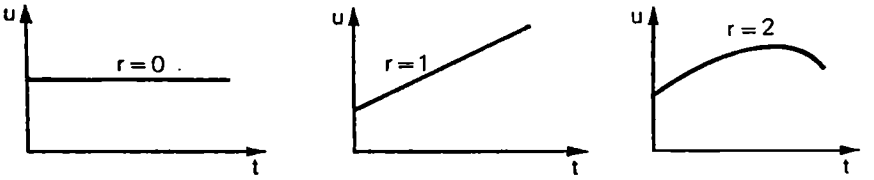


Fig. 10.13. Polynomial signals for various r

on, whose forms are shown in Fig. 10.13. We assume also that at discrete time instants the interference may be represented by white noise (see Example 5.19) with zero mean and variance σ^2.

10.6.6 Filtering by Least Squares

We find the derivatives of the signal by differentiating (10.92) r times with respect to τ

$$\dot{u}(t-\tau) = \dot{u}(t) - \ddot{u}(t) + \ldots + \frac{(-\tau)^{r-1}}{(r-1)!}u^{(r)}(t)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$u^{(r)}(t-\tau) = u^{(r)}(t)$$

Filtering algorithms are implemented in a form of computer codes. To convert these algorithms to a discrete form suitable for computers the time variable is quantized by letting $t_n = nT_0$ with $n = 0, 1, 2$, etc., and $\tau_i = iT_0$ with $i = 0, 1, 2$, etc., where T_0 is the quantization period usually taken to be a unit of time.

The discrete model of the signal may be represented by the system of difference equations

$$u_{n-i} = u_n - i\dot{u}_n + \frac{i^2}{2}\ddot{u}_n + \ldots + \frac{(-i)^r}{r!}u_n^{(r)}$$

$$\dot{u}_{n-1} = \dot{u}_n - i\ddot{u}_n + \ldots + \frac{(-i)^{r-1}}{(r-1)!}u_n^{(r)}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$u_{n-i}^{(r)} = u_n^{(r)}$$

or in vector form

$$\mathbf{x}_{n-i} = \Phi_i \mathbf{x}_n \tag{10.95}$$

where $\mathbf{x}_n = (u_n, \dot{u}_n, \ldots, \overset{(r)}{u}_n)^{\mathrm{T}}$ and

$$\Phi_i = \begin{Vmatrix} 1 & -i & \frac{i^2}{2} & \ldots & \frac{(-i)^r}{r!} \\ 0 & 1 & -i & \ldots & \frac{(-i)^{r-1}}{(r-1)!} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \end{Vmatrix} \tag{10.96}$$

We wish to obtain an estimate $\hat{\mathbf{x}}_n$ by using the last l observations of a signal including the current observation. These observations may be represented in the form

$$z_{n-i} = u_{n-i} + s_{n-i}, \quad i = 0, 1, \ldots, l-1 \tag{10.97}$$

For technical convenience of writing (10.97) we introduce the vector

$$\mathbf{h}_i = \left(1, -i, \frac{i^2}{2}, \ldots, \frac{(-i)^r}{r!}\right) \tag{10.98}$$

being the first row in the transformation matrix Φ_i. If we note that $\mathbf{h}_0 = (1, 0, \ldots, 0)$, then for $\mathbf{h}_i$ we have

$$\mathbf{h}_i = \mathbf{h}_0 \Phi_i \tag{10.99}$$

Using the vectors $\mathbf{h}_0$ and $\mathbf{h}_i$, the first component u_{n-i} of $\mathbf{x}_{n-i}$ may be represented as

$$\mathbf{u}_{n-i} = \mathbf{h}_0 \mathbf{x}_{n-i} = \mathbf{h}_0 \Phi_i \mathbf{x}_n = \mathbf{h}_i \mathbf{x}_n \tag{10.100}$$

The observations (10.97) can be written as

$$z_{n-i} = \mathbf{h}_i \mathbf{x}_n + s_{n-i}, \; i = 0, 1, \ldots, l-1 \tag{10.101}$$

or

$$\begin{aligned} z_{n-(l-1)} &= \mathbf{h}_{l-1} \mathbf{x}_n + s_{n-(l-1)} \\ z_{n-(l-2)} &= \mathbf{h}_{l-2} \mathbf{x}_n + s_{n-(l-2)} \\ &\cdots\cdots\cdots\cdots \\ z_n &= \mathbf{h}_0 \mathbf{x}_n + s_n \end{aligned} \tag{10.102}$$

These l results can be represented as one observation vector

$$\mathbf{Z}_{n,l} = \mathbf{H}_l \mathbf{x}_n + \mathbf{n}_{n,l} \tag{10.103}$$

where

$$\mathbf{Z}_{n,l} = \begin{pmatrix} z_{n-(l-1)} \\ z_{n-(l-2)} \\ \vdots \\ z_n \end{pmatrix}, \quad \mathbf{H}_l = \begin{pmatrix} \mathbf{h}_{l-1} \\ \mathbf{h}_{l-2} \\ \vdots \\ \mathbf{h}_0 \end{pmatrix}, \quad \mathbf{n}_{n,l} = \begin{pmatrix} s_{n-(l-1)} \\ s_{n-(l-2)} \\ \vdots \\ s_n \end{pmatrix}$$

Equation (10.103) has the same form as (10.88). Recalling that $\mathbf{n}_{n,l}$ is the vector of uncorrelated interference, an optimal least squares estimate will be evaluated by (10.91) which can be conveniently written as two relations by denoting

$$P_l = \mathbf{H}_l^T \mathbf{H}_l \tag{10.104}$$

so that

$$\hat{\mathbf{x}} = \mathbf{P}_l^{-1} \mathbf{H}_l^T \mathbf{Z}_{n,l} \tag{10.105}$$

These relations can be readily implemented on a computer in matrix form. For this purpose the operator loads into the computer memory the matrix of coefficients $\mathbf{H}_l$ and the array of observational data $\mathbf{Z}_{n,l}$. The resulting estimate $\hat{\mathbf{x}}_n$ is obtained by the finite set of observations available in the computer storage. Digital filters operating by the relations (10.104) and (10.105) are referred to as *finite memory structures.*

These filters allow current estimates $\hat{\mathbf{x}}_n$ to be obtained for variable n. For every n the computer stores l previous observations of the process. Therefore, the set of data stored by the computer will vary with n. At each new step one new observation z_{n+1} is added to the set $\mathbf{Z}_{n,l}$ and the earliest sample $z_{n-(l-1)}$ is discarded. This process of updating in obtaining current estimates is sometimes referred to as a *sliding window.*

In on-line filtering, if $n \leqslant i$, it is assumed $z_{n-i} = 0$. Since the observation process starts with $n = 1$, then at $n = 1, 2, \ldots, l - 1$ the set $\mathbf{Z}_{n,l}$ will not contain a complete set of samples and consequently cannot yield the estimate $\hat{\mathbf{x}}_n$. Therefore, the first $l - 1$ observations will represent the transient process of the digital filter.

Sometimes it is required to obtain a vector of estimates not at time n, but at time $n + m$ that is a future instant for $m > 0$ or a past instant for $m < 0$ with respect to the present discrete time n. An estimate $\hat{\mathbf{x}}_{n+m}$ will be a prediction of $\hat{\mathbf{x}}_n$ for m steps ahead or an *interpolation* of $\hat{\mathbf{x}}_n$ m steps backwards. Having a vector of estimates $\hat{\mathbf{x}}_n$ and using (10.95 , we get

$$\hat{\mathbf{x}}_{n+m} = \mathbf{\Phi}_{-m} \hat{\mathbf{x}}_n \tag{10.106}$$

10.6.7 Weighting Coefficient Representation

At times it is desirable to use the vector of weight coefficients

$$\mathbf{k}(l, m) = \begin{pmatrix} k_0(l, m) \\ \cdot \cdot \cdot \cdot \cdot \\ k_r(l, m) \end{pmatrix} \tag{10.107}$$

to obtain an estimate of a signal state

$$\hat{\mathbf{x}}_n = \sum_{m=0}^{l-1} \mathbf{k}(l,\ m)\, z_{n-m} \tag{10.108}$$

or in expanded form

$$\hat{u}_n^{(j)} = \sum_{m=0}^{l-1} k_j(l,\ m)\, z_{n-m}, \quad j = 0,\ \ldots,\ r \tag{10.109}$$

To reduce (10.105) to (10.108) we develop the product

$$\mathbf{H}_l^{\mathrm{T}} \mathbf{Z}_{n,l} = (\mathbf{h}_{l-1}^{\mathrm{T}},\ \ldots,\ \mathbf{h}_0^{\mathrm{T}}) \begin{pmatrix} z_{n-(l-1)} \\ \vdots \\ z_n \end{pmatrix} = \sum_{m=0}^{l-1} \mathbf{h}_m^{\mathrm{T}} z_{n-m}$$

so that (10.105) becomes

$$\hat{\mathbf{x}}_n = \mathbf{P}_l^{-1} \sum_{m=0}^{l-1} \mathbf{h}_m^{\mathrm{T}} z_{n-m} \tag{10.110}$$

Comparison with (10.108) yields

$$\mathbf{k}(l,\ m) = \mathbf{P}_l^{-1}\, \mathbf{h}_m^{\mathrm{T}} \tag{10.111}$$

At $r = 0$ (the case of constant signal $u(t)$), we obtain $\mathbf{h}_m = 1$, $\mathbf{H}_l = (1,\ \ldots,\ 1)^{\mathrm{T}}$, $\mathbf{P}_l = l$, $\mathbf{P}_l^{-1} = 1/l$. Consequently,

$$\mathbf{k}(l,\ m) = k_0(l,\ m) = 1/l$$

By virtue of (10.109) we obtain

$$\hat{u}_n = \frac{1}{l} \sum_{m=0}^{l-1} z_{n-m} \tag{10.112}$$

We leave it as an exercise to the reader to verify that for a linearly variable signal (ramp) of $r = 1$ the estimates $\hat{u}_n$ and $\hat{\dot{u}}_n$ are found by the formulas

$$\begin{aligned} \hat{u}_n &= \sum_{m=0}^{l-1} k_0(l,\ m)\, z_{n-m} \\ \hat{\dot{u}}_n &= \sum_{m=0}^{l-1} k_1(l,\ m)\, z_{n-m} \end{aligned} \tag{10.113}$$

where

$$\begin{aligned} k_0(l,\ m) &= 2(2l - 1 - 3m)/[l(l+1)] \\ k_1(l,\ m) &= 6(l - 1 - 2m)/[(l-1)l(l+1)] \end{aligned} \tag{10.114}$$

Prediction is then performed by (10.106) as

$$\hat{u}_{n+m} = \hat{u}_n + m\dot{\hat{u}}_n, \quad \dot{\hat{u}}_{n+m} = \dot{\hat{u}}_n \tag{10.115}$$

Example 10.16. Let $r = 1$ and $l = 5$. The estimates obtained by (10.114) take the form

$$\hat{u}_n = (1/5)\,(3z_n + 2z_{n-1} + z_{n-2} - z_{n-4})$$

$$\dot{\hat{u}}_n = (1/10)\,(2z_n + z_{n-1} - z_{n-3} - 2z_{n-4})$$

To achieve the desired performance of a digital filter the designer should match the parameters of real signals and interferences to be handled with such filter parameters as the sampling period, the degree r of the polynomial representing the signal, and the span of the observation interval $T = lT_0$ which defines the number of observations l stored in the filter memory. All these topics are discussed in depth in Ref. [67].

10.6.8 The Kalman Filter

Fixed memory filters have a number of disadvantages limiting their practical use. These disadvantages follow from (10.108) and are as follows.

1. Finite memory structures are hard to implement as

(a) they require that the last l observation data should be stored and updated, for on-line filtering, as new observations are arrived;

(b) the filter memory should store l weighting coefficients $\mathbf{k}(l, m)$, $m = 0, \ldots, l-1$, computed in advance;

(c) after each observation all the set of stored data should be recomputed.

Therefore on-line filtering requires large memory capacity and fast processors to perform all the computations in the interval between two adjacent samplings.

2. The vector of estimates $\mathbf{x}_n$ cannot be computed until all the data of l first observations are stored. Therefore this type of filter has a long transition process which is not always affordable as the information about the process is desired as early as possible.

The last of the above disadvantages can be alleviated by employing structures with increasing memory.

In such expandable memory filters, the number of discrete states l of an observed signal used to determine $\hat{\mathbf{x}}_n$ should coincide with the number n of the present input. Therefore, the vector of estimates $\hat{\mathbf{x}}_n$ for an extended memory filter can be obtained from the relation

(10.108) for finite memory filters by replacing l with n, namely,

$$\hat{\mathbf{x}}_n = \sum_{m=0}^{n-1} \mathbf{k}(n, m)\, \mathbf{z}_{n-m} \tag{10.116}$$

With this relation current estimates $\hat{\mathbf{x}}_n$ may be obtained beginning with $n = r + 1$, i.e. beginning with the number of observations equal to the number of components in $\mathbf{x}_n$. Specifically, at $r = 0$ such a filter will exhibit a zero span of the transition process. As n increases the smoothing properties of the filter are enhanced, and the accuracy of estimation is improved.

However, a realization of such a filter by the formula (10.116) would not be practical as it involves a cumbersome computational procedure for the coefficients $\mathbf{k}(n, m)$ for $m = 0, \ldots, n - 1$, which will vary with each new observation. Moreover, the observations increasing in number with n should be stored in filter memory. Therefore, structure with increasing memory should be realized as *recursive*, i.e. involving feedback connections, structures.

Suppose that by time n we have obtained an estimate $\hat{\mathbf{x}}_{n-1}$ containing all information about the preceding observations $z_1, \ldots, z_{n-1}$. The estimate $\hat{\mathbf{x}}_n$ will be obtained after a sequential observation z_n has been taken. A natural question arises as to whether or not we can obtain this estimate without direct use of the data $z_1, \ldots, z_{n-1}$ but employing instead only the information contained in the estimate $\hat{\mathbf{x}}_{n-1}$. This procedure is feasible and it became known as *recursive filtering*. In general terms it is as follows.

(*i*) Propagate the estimate $\hat{\mathbf{x}}_{n-1}$ one step forward using (10.106) and determine the *predicted*, or *a priori*, *estimate*

$$\mathbf{x}_n^p = \Phi_{-1}\hat{\mathbf{x}}_{n-1} \tag{10.117}$$

(*ii*) Use the *present input* z_n to convert this *a priori* estimate into the true, i.e. *a posteriori*, estimate $\hat{\mathbf{x}}_n$.

This procedure is repeated at each step beginning with $n = r + 2$. The initial condition for this procedure is provided by the estimate $\hat{\mathbf{x}}_{r+1}$ derived by the formulas (10.108) for finite memory filter at $n = r + 1$.

To obtain an algorithm of recursive filtering we consider observation k

$$z_k = u_k + s_k = \mathbf{h}_0\mathbf{x}_k + s_k$$

Observing that $\mathbf{x}_k = \mathbf{x}_{n-(n-k)} = \Phi_{n-k}\mathbf{x}_n$, we get

$$z_k = \mathbf{h}_0\Phi_{n-k}\mathbf{x}_n + s_n = \mathbf{h}_{n-k}\mathbf{x}_n + s_k$$

Letting $k = 1, 2, \ldots, n-1, n$, we obtain the description of the complete set of observations in the first n steps

$$\begin{aligned} z_1 &= \mathbf{h}_{n-1}\mathbf{x}_n + s_1 \\ z_2 &= \mathbf{h}_{n-2}\mathbf{x}_n + s_2 \\ &\cdots\cdots\cdots \\ z_{n-1} &= \mathbf{h}_1\mathbf{x}_n + s_{n-1} \\ z_n &= \mathbf{h}_0\mathbf{x}_n + s_n \end{aligned} \tag{10.118}$$

Consider the first $n-1$ observations. In vector form they are

$$\mathbf{Z}_{n-1} = \mathbf{H}_{n-1}\mathbf{x}_n + \mathbf{n}_{n-1}$$

where $\mathbf{Z}_{n-1} = (z_1, \ldots, z_{n-1})^{\mathrm{T}}$, $\mathbf{H}_{n-1} = (\mathbf{h}_{n-1}, \ldots, \mathbf{h}_1)^{\mathrm{T}}$, and $\mathbf{n}_{n-1} = (s_1, \ldots, s_{n-1})^{\mathrm{T}}$.

The solution of this set of equations yields the predicted estimate $\mathbf{x}_n^p$ because the set of observations does not include z_n. Utilizing (10.91), we obtain

$$\mathbf{x}_n^p = \mathbf{P}_{n-1}^{-1}\mathbf{H}_{n-1}^{\mathrm{T}}\mathbf{Z}_{n-1}, \quad \mathbf{P}_{n-1} = \mathbf{H}_{n-1}^{\mathrm{T}}\mathbf{H}_{n-1} \tag{10.119}$$

The full set of n observations is described by the system $\mathbf{Z}_n = \mathbf{H}_n\mathbf{x}_n + \mathbf{n}_n$. Its solution yields already the posterior (current) estimate $\hat{\mathbf{x}}_n$. We use (10.91) to represent it in the form

$$\hat{\mathbf{x}}_n = \mathbf{P}_n^{-1}\mathbf{H}_n^{\mathrm{T}}\mathbf{Z}_n, \quad \mathbf{P}_n = \mathbf{H}_n^{\mathrm{T}}\mathbf{H}_n \tag{10.120}$$

Now, we express the posterior estimate $\hat{\mathbf{x}}_n$ through the prior estimate $\mathbf{x}_n^p$. For this purpose we write the vectors $\mathbf{Z}_n = (\mathbf{Z}_{n-1}, z_n)^{\mathrm{T}}$ and $\mathbf{H}_n = (\mathbf{H}_{n-1}, \mathbf{h}_0)^{\mathrm{T}}$. Substituting these vectors in (10.120), we get

$$\hat{\mathbf{x}}_n = \mathbf{P}_n^{-1}(\mathbf{H}_{n-1}^{\mathrm{T}}, \mathbf{h}_0^{\mathrm{T}})\begin{pmatrix}\mathbf{Z}_{n-1}\\ z_n\end{pmatrix} = \mathbf{P}_n^{-1}(\mathbf{H}_{n-1}^{\mathrm{T}}\mathbf{Z}_{n-1} + \mathbf{h}_0^{\mathrm{T}}z_n)$$

Also,

$$\hat{\mathbf{x}}_n - \mathbf{x}_n^p = \hat{\mathbf{x}}_n - \mathbf{P}_n^{-1}\mathbf{P}_n\mathbf{x}_n^p = \mathbf{P}_n^{-1}(\mathbf{H}_{n-1}^{\mathrm{T}}\mathbf{Z}_{n-1} + \mathbf{h}_0^{\mathrm{T}}z_n - \mathbf{P}\mathbf{x}_n^p)$$

From (10.119) it follows $\mathbf{H}_{n-1}^{\mathrm{T}}\mathbf{Z}_{n-1} = \mathbf{P}_{n-1}\mathbf{x}_n^p$ so that

$$\hat{\mathbf{x}}_n - \mathbf{x}_n^p = \mathbf{P}_n^{-1}[\mathbf{h}_0^{\mathrm{T}}z_n - (\mathbf{P}_n - \mathbf{P}_{n-1})\mathbf{x}_n^p] \tag{10.121}$$

If we note that

$$\mathbf{P}_n = \mathbf{H}_n^{\mathrm{T}}\mathbf{H}_n = (\mathbf{H}_{n-1}^{\mathrm{T}}, \mathbf{h}_0^{\mathrm{T}})\begin{pmatrix}\mathbf{H}_{n-1}\\ \mathbf{h}_0\end{pmatrix} = \mathbf{H}_{n-1}^{\mathrm{T}}\mathbf{H}_{n-1} + \mathbf{h}_0^{\mathrm{T}}\mathbf{h}_0 = \mathbf{P}_{n-1} + \mathbf{h}_0^{\mathrm{T}}\mathbf{h}_0$$

then

$$\mathbf{P}_n - \mathbf{P}_{n-1} = \mathbf{h}_0^{\mathrm{T}}\mathbf{h}_0$$

Substituting this difference into (10.121), we arrive at the desired relation

$$\hat{\mathbf{x}}_n = \mathbf{x}_n^p + \mathbf{P}_n^{-1}\mathbf{h}_0^T (z_n - \mathbf{h}_0\mathbf{x}_n^p) \tag{10.122}$$

The expanded memory filter defined by (10.122) is a particular case of a filtering algorithm known as the *Kalman filter* [66, 68]. The structure of this filter and the relevant computational steps are

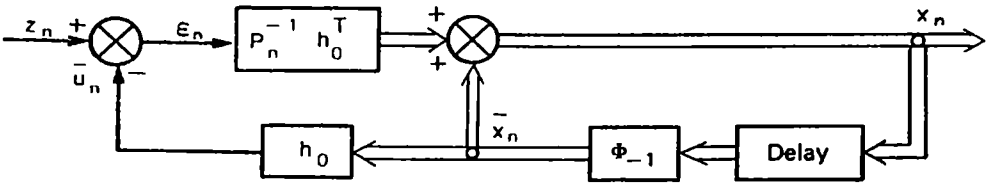


Fig. 10.14. Block diagram of a Kalman filter

represented in the block diagram in Fig. 10.14. Double-shafted lines indicate connections through which the information flow is in vector form.

In practical applications of the algorithm (10.122) it should be kept in mind that the respective predicted estimate $\mathbf{x}_n^p$ is obtained by (10.117) and the quantity $\mathbf{h}_0\mathbf{x}_n^p$ is the first component of vector $\mathbf{x}_n^p$.

Specific forms of filter algorithms for constant and ramp signals are as follows:

(a) for $r = 0$ (the initial conditions $\hat{u}_1 = z_1$)

$$\hat{u}_n = \hat{u}_{n-1} + (z_n - \hat{u}_{n-1})/n, \; n = 2, 3, \text{ etc.}$$

(b) for $r = 1$ (the initial conditions $\hat{u}_2 = z_2$, $\dot{\hat{u}}_2 = z_2 - z_1$)

$$\hat{u}_n = u_{n-1}^p + 2(2n-1)(z_n - u_n^p)/n(n+1)$$

$$\dot{\hat{u}}_n = \dot{u}_n^p + 6(z_n - u_n^p)/n(n+1)$$

where $u_n^p = \hat{u}_{n-1} + \dot{\hat{u}}_{n-1}$, $\dot{u}_n^p = \dot{\hat{u}}_{n-1}$, $n = 3, 4$, etc.

10.6.9 Exponential Filters

To provide better insight into an important feature of expanded memory filters we revisit Eq. (10.122) for a more detailed consideration. In this expression, the vector of estimates of signal states $\hat{\mathbf{x}}_n$ is obtained as a sum of the vector of predicted estimates $\mathbf{x}_n^p$ and a me-

asured residual (or error) $z_n - \mathbf{u}_n^p$ taken with weighting coefficients defined by the components of the vector $\mathbf{P}_n^{-1}\mathbf{h}_0^T$. At small n when the prediction accuracy is not very high, the corrections introduced by the measured residual by the present input z_n will be considerable since the components of $\mathbf{P}_n^{-1}\mathbf{h}_0^T$ are appreciably far from zero. However, as n increases these weighting coefficients decrease and tend to zero as $n \to \infty$, which may be easy to verify by considering the above examples of Kalman filter. This circumstance reflects the fact that as n increases, i.e. the sampling interval of signal observation expands, a more efficient smoothing of noise takes place and the accuracy of the predicted estimate $\mathbf{x}_n^p$ improves. It should be recognized, however, that such a growth of accuracy with n will take place only when the actual signal has the same form as the model of signal adopted for prediction.

The predicted estimate $\mathbf{x}_n^p$ is obtained by (10.117) for the polynomial model (10.92). At the same time, actual signals may be represented by this model on limited time intervals only. Therefore, as n increases the actual signal deviated ever more from the polynomial model and the prediction becomes ever more inaccurate.

This situation can be corrected by raising the weight of the present input z_n at large n thus correcting the inaccuracy of the predicted estimate $\mathbf{x}_n^p$. This can be achieved by imposing a restriction on excessive drop of the weight $\mathbf{P}_n^{-1}\mathbf{h}_0^T$ as n increases. To achieve this goal we set up a certain number of observations l at which the predicted estimate $\mathbf{x}_n^p$ is still sufficiently accurate and will employ the expanded memory filter (10.122) to obtain the current estimate $\hat{\mathbf{x}}_n$ only for $n \leqslant l$. In this setting, the weight $\mathbf{P}_n^{-1}\mathbf{h}_0^T$ will gradually decrease attaining the value of $\mathbf{P}_l^{-1}\mathbf{H}_0^T$ at $n = l$.

Beginning from time $n = l$ we prohibit a further decrease of $\mathbf{P}_n^{-1}\mathbf{h}_0^T$ keeping it at the level $\mathbf{P}_l^{-1}\mathbf{h}_0^T$. This is a tradeoff between the acceptable accuracy of prediction and the correction of the measured residual by the present input signal. The resulting filtering algorithm for $n \geqslant l$ takes the form

$$\hat{\mathbf{x}}_n = \mathbf{x}_n^p + \mathbf{P}_l^{-1}\mathbf{h}_0^T (z_n - u_n^p) \qquad (10.123)$$

Practical structures implementing this algorithm are sometimes referred to as *exponential filters*. They are in wide use for obtaining current estimates of signals that allow approximation by polynomials of zero ($r = 0$) and first ($r = 1$) order on finite-length intervals lT_0.

At $r = 0$, the exponential filter, known as the α-filter or exponential smoothing structure, is described by

$$\hat{u}_n = \hat{u}_{n-1} + \alpha (z_n - \hat{u}_{n-1}), \ \alpha = 1/l \qquad (10.124)$$

At $r = 1$, the estimates of the exponential filter, called now an α- or β-filter, are obtained by the relations

$$\hat{u}_n = \dot{u}_n^p + \alpha (z_n - u_n^p)$$
$$\dot{\hat{u}}_n = \dot{u}_n^p + \beta (z_n - u_n^p) \tag{10.125}$$

where $u_n^p = \hat{u}_{n-1} + \dot{\hat{u}}_{n-1}$, $\dot{u}_n^p = \dot{\hat{u}}_{n-1}$, $\alpha = 2(2l - 1)/l(l + 1)$, and $\beta = 6/l(l + 1)$.

EXERCISES

10.1. Use the data of Table 10.2 to find a strategy based on the minimax principle of additional losses in the processing line problem.

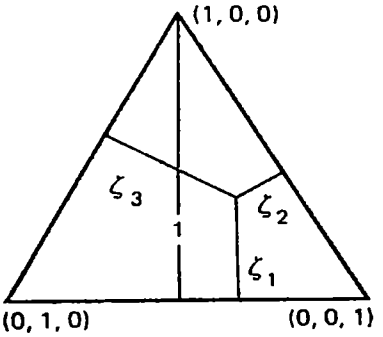


Fig. 10.15. Space of mixed strategies for the case of three states of nature

10.2. With reference to Fig. 10.6 calculate the minimax and Bayes strategies in the sample information problem of processing line.

10.3. A message u corrupted by a noise s of a gaussian distribution $N(0, \sigma^2)$ is received as $z = u + s$. The message can take on two values $u = 0$ (no signal, ϑ_1) and $u = u_0 = \text{const}$ (ϑ_2). The receiver has a threshold z_0 and decides that a message u_0 was sent if $z > z_0$. Determine the threshold z_0 using the solution techniques for the binary decision problem and letting $\zeta = 0.25$ and $w = 2$.

10.4. Demonstrate that in the case of three states of nature $\Theta = (\vartheta_1, \vartheta_2, \vartheta_3)$ the possible mixed strategies of the statistician $\xi(\vartheta) = (\zeta_1, \zeta_2, \zeta_3)$ form a space that has a form of a triangle with unit altitude, as shown in Fig. 10.15.

11

SCHEDULING AND COMBINATORIAL PROBLEMS. THEORY OF QUEUES

11.1 Scheduling and Sequencing

11.1.1 General

Many problems of planning, management and social services are concerned with optimal sequencing of tasks or jobs subject to certain constraints on the ordering of operations, duration of each job, resources allocated to the project, and the like [73]. Such problems are many and diverse and treated by many disciplines depending on the characteristic features. This section will be devoted to the problems involving the evaluation of the ordering and schedule of performing large projects that consist of many connected jobs or operations requiring numerous contractors to do and associated with considerable costs

The companion topic of this chapter is the theory of queues that has to do with assigning priorities in serving customers arriving at servicing facilities which may be clerks, channels or routes, and machines or regulations. Similarity of the models used to solve both types of problems should justify their coverage in the framework of one chapter.

11.1.2 Problem Formulation

Any targetted activity may be viewed as a time-varying process aimed at the realization of a certain set of operations (in a job) or jobs (in a project), called a *program*

$$S = \{P_1, \ldots, P_s\} \tag{11.1}$$

Frequently a number of constraints are imposed on the fulfilment of such projects, and these constraints are of two kinds [37].

Constraints α. *Schedule criteria*. These constraints describe the mutual dependence of jobs in a sequence. Every job $P_i \in S$ may be put in correspondence to a set $\Pi_i \subset S$ of the immediately preceding jobs which must be completed before a job P_i can be started, and also some jobs that do not belong to S and may be viewed as external factors necessary to perform P_i. For example, the erection of walls in a building cannot be started until a foundation is laid (job from S) and unless the brick and mortar are supplied to the construction site.

The preparation of mortar and brick delivery may not enter the list of jobs scheduled for the project.

To describe some other conditions entering constraints the development of the project is convenient to discuss as proceeding in discrete time by assuming some time interval Δt to be a unit of time. This interval may be called a conditional day or year, but we shall call it a step.

Some operations can be completely finished in one step, others in a few steps. Constraints α should include a condition of stating whether or not a particular job can be interrupted (preempted).

Let t be the number of a step ($t = 1, 2$, etc.), and $y_i(t)$ be a portion of job P_i made at step t. We shall recognize this portion of P_i as a unit job. Denote by $S(t)$ the list of jobs made at step t.

Assume that job P_i consists of m unit jobs y_i and no preemption of this job is allowed. The inadmissibility of preemption implies that if a job begins at step t, then its unit portions y_i should enter the lists of jobs $S(t)$, $S(t+1)$, . . ., $S(t+m-1)$.

Constraints β. *Process constraints.* These constraints are imposed on the amount of resources allocated to the project S.

Denote by $Q(t)$ the vector of resource that can be allocated to the job at step t. The components of $Q(t)$ may be as follows, Q_1, raw materials; Q_2, machinery; Q_3, manpower; etc. Define also $q[y_i(t)]$ to be the amount of resource required to fulfill a unit job $y_i(t)$. In this notation, constraint β may be written as

$$\sum_{i-1}^{s} q[y_i(t)] \leqslant Q(t), \quad t = 1, 2, \ldots, M \tag{11.2}$$

where M is the number of steps in the sequence.

It is also assumed that a unit job $y_i(t)$ is completed if it has been provided with all the necessary resources, or is not attempted at all if some resource is lacking. The general situation with the proportion of job done independent of the amount of resource needs more complicated models to handle.

A typical scheduling problem seeks to compile for each step t the list of jobs $S(t) = \{P_{t1}, \ldots, P_{tl}\}$ that must be performed at this step, and the proportion $y_i(t)$ of each of these jobs subject to the constraints imposed on the order discipline, and allocated resources.

If the number of jobs is large, the number of schedule versions may be enormous. Choice of a specific version is due to the project *lead time.* If the list of jobs and resources is specified, then a best (optimal) version of schedule will be such that minimizes the mission time which is called then the *lead time.*

11.1.3 Types of Problems

Let a program of jobs $S = \{P_1, \ldots, P_s\}$ be given subject to constraints α dictating the sequence of jobs and constraints β restricting the allocated resources. In practical situations constraints α connect not all jobs but act only within some independent groups of jobs, the groups being not tied by strict restrictions.

Under the circumstances, the program S may be represented as a partition $M = \{R_1, \ldots, R_m\}$ whose elements R represent individual subprograms not related by any priority rules, so that each subprogram R may be performed almost independently of other subprograms. However, within each subprogram $R_i = \{P_{i1}, \ldots, P_{ir}\}$ the sequence of jobs is subject to strict constraints.

Thus, a scheduling case may be broken down into two independent problems.

In the first type of problem, strict constraints α predetermine, to a considerable degree, the order of individual jobs so that the number of possible versions to choose from is here not very large. Some versions are always possible and among them there are some with minimal mission time, more rational resource allocation than the others, and the like. These problems are handled within the framework of an approach known as network planning and control.

The second type of problems have only insignificant constraints on the sequencing of jobs. This does not mean, however, that they are indifferent to the order of jobs as the latter dictates the allocation of resources to steps. A transition from one job to another takes time, may call for readjustment of machinery or replacement, changeover of works and such. Therefore, here again we seek a schedule that would be best in a certain sense. However, absence of constraints in this type of problem requires investigation of very large number of combinations of job sequences and assessment of each combination. Accordingly, such problems are called *combinatorial* or *ordering problems.*

11.2 Project Planning and Control

11.2.1 The Notion of Network

Graphical methods are in wide use for rational planning of large projects including the development of a set of interconnected jobs. These may be the construction of a large and complicated structure, the development of a new system or product, a large servicing or reconstruction project, manufacture and assembly of large structures such as aircraft, ships, space vehicles, etc. These methods represent the set of jobs in the form of an oriented graph whose arcs represent

jobs and the vertices represent some events such as completion of individual jobs. The sequence of arcs in such a graph defines the order in which the jobs are to be made.

Figure 11.1 shows an example of a network scheduling the jobs involved in the development of a device. This representation makes a complicated project more tractable and allows a better allocation of the available labor and material resources.

The first step in constructing a network is to partition the project into individual jobs. Each job takes time to do, therefore has its start

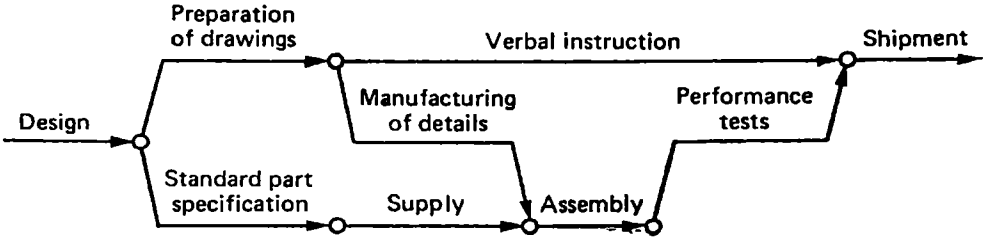


Fig. 11.1. Network diagram of a device development project

and finish. These instants are usually easy to determine. Jobs without definite end points should not be introduced into the list and are subject to further refinement.

The list of jobs is usually compiled by experts in the particular project. It should be recognized, however, that even an expert is not guaranteed from mistakes, such as overlooking of some jobs. Therefore, the list of jobs should be agreed with the contractors who may easily trace out the missing jobs and other drawbacks of the list.

The compilation of a list of jobs is paralleled by the development of constraints such as the duration of each job, cost of each job, manpower loading, and the list of preceding jobs necessary to initiate a successor job. All these data are arrayed into a table such as Table 11.1 for 16 jobs.

11.2.2 Ranging of Jobs

The order of jobs shown in Table 11.1 is to a certain extent random in character. It is not evident which of these jobs are important, i.e. not clear which jobs deserve more attention of the entire project.

To eliminate this drawback we assign to each job some weight reflecting its importance and, thus, dictating, to a certain degree, the order in which they should be performed. Assignment of weights results in ordering or ranging of the jobs.

Of the many methods to range a set of jobs we take a simple and natural one [73]. We assign to each job a weight equal to the sum of the following jobs and their weights. The jobs of end arcs receive a

Table 11.1

Job type	Months	Cost tsd rbl	Manpower, workers/month	Preceding jobs
A	1	2.0	8	—
B	4	13.0	13	*A*, *C*, *D*
C	1	5.0	20	—
D	1	5.0	20	—
E	2	2.0	4	—
F	3	12.0	16	*D*, *E*
G	4	5.0	5	*F*
H	1	1.0	4	*C*
I	5	4.0	3	*B*, *G*, *N*
K	2	1.0	2	*C*
L	2	3.0	6	*H*
M	2	1.0	2	*H*
N	4	0.5	1	*A*, *D*
P	2	5.0	10	*F*
Q	6	15.0	10	—
R	3	5.0	7	*Q*

relative weight corresponding to their importance. We shall assume that all terminal jobs are equally important and assign to them identical weight equal to unity.

Table 11.2

Jobs and constraints	Weight of job	Jobs and constraints	Weight of job
$A < B, N$	$2+2+2=6$	I —	1
$B < I$	$1+1=2$	K —	1
$C < B, H, K$	$3+2+4+1=10$	L —	1
$D < B, F, N$	$3+2+5+2=12$	M —	1
$E < F$	$1+5=6$	$N < I$	$1+1=2$
$F < G, P$	$2+2+1=5$	P —	1
$G < I$	$1+1=2$	$Q < R$	$1+1=2$
$H < L, M$	$2+1+1=4$	R —	1

The assignment procedure is illustrated by Table 11.2. The left column shows the jobs and those following them. These data may be borrowed from Table 11.1. The last column of this table in the second row shows that job B is preceded by A, C, and D, which may be written as $A < B$, $C < B$, and $D < B$. Hence, job B immediately fol-

lows A, C, and D which fact is noted in the respective rows of Table 11.2.

The weight is calculated by inspecting the left column of Table 11.2 from the bottom upwards. This inspection has to be made several times. At the beginning we assign weight to end jobs R, P, M, L, K and I. In every next inspection we calculate the weight of those jobs for which the weight of the following jobs is already known.

Table 11.3

Job	Men	τ	ω'_k	ω''_k	t'_k	t''_k	$\Delta\tau$	t^0
D	20	1	—	B, N, F	1	2	1	0
C	20	1	—	B, H, K	1	5	4	0
E	4	2	—	F	2	2	0	0
A	8	1	—	B, N	1	5	4	0
F	16	3	D, E	G, P	5	5	0	2
H	4	1	C	L, M	2	12	10	1
Q	10	6	—	R	6	11	5	0
N	1	4	A, D	I	5	9	4	1
G	5	4	F	I	9	9	0	5
B	13	4	A, C, D	I	5	9	4	1
R	7	3	Q	—	9	14	5	6
P	10	2	F	—	7	14	7	5
M	2	2	H	—	4	14	10	2
L	6	2	H	—	4	14	10	2
K	2	2	C	—	3	14	11	1
I	3	5	B, G, N	—	14	14	0	9

Note: τ = job time; ω'_k and ω''_k = preceding and following jobs; t'_k and t''_k = early and latest termination terms.

The ordered set of jobs can be obtained by rearranging jobs in Table 11.3 in the order of decreasing weight. For jobs with identical weight the order may be chosen at will. We also write into this table the data on the rate labor (requirement) of jobs, their scheduled time τ, the sets of preceding and following jobs, ω', and ω'', and the terms of start, finish, and slack time (float) for each job. The evaluation of the latter quantities will be discussed below.

11.2.3 Slack Time (Float) Evaluation

For some jobs the onset may be varied within slack time (float) without having to disturb the required sequence of jobs and the order time of the whole mission. Denote by τ_h the time required to perform

job k, and by t'_k and t''_k the earliest and latest time of completion. Then the float for job k is

$$\Delta\tau_k = t''_k - t'_k \tag{11.3}$$

The mission-time minimization problem reduces thereby to the evaluation of instants t'_k and t''_k for each job k.

Evaluation of t'_k. Consider a job $i \in \omega'_k$, i.e. one of the jobs immediately preceding job k. If t'_k is the end instant of job k, then it starts at $t'_k - \tau_k$. Since job k cannot start until all preceding jobs have been completed, then the condition must hold

$$t'_k - \tau_k \geqslant t'_i, \; i \in \omega'_k \tag{11.4}$$

The equality in this relationship defines the earliest finish time of job k

$$t'_k = \max_{i \in \omega'_k} (t'_i + \tau_k) \tag{11.5}$$

Therefore, to know t'_k we should know t'_i for all immediately preceding jobs. Accordingly, the evaluation of t'_k should begin with the jobs that have no predecessors, i.e. head the list, and for which the earliest time of completion means the duration of the job.

The greatest time t'_k means the mission time for the entire project,

$$T = \max_k t'_k \tag{11.6}$$

In the case under consideration $T = 14$ months.

Evaluation of t''_k. Consider a job $j \in \omega''_k$, i.e. one of jobs immediately following job k. If t''_i is its latest finish time, then the latest starting time for it will be $t''_j - \tau_j$. Since job j cannot begin until job k has been finished, then

$$t''_k \leqslant t''_j - \tau_j, \quad j \in \omega''_k \tag{11.7}$$

whence it follows that

$$t''_k = \min_{j \in \omega''_k} (t''_j - \tau_j) \tag{11.8}$$

Because to determine t''_k we need to know the instants when the immediately following jobs terminate, the consideration should begin with jobs followed by no other jobs, i.e. with I, K, L, M, P, and R. For these jobs, the latest finish time is taken as the mission time T.

Once the times t'_k and t''_k have been determined for every job, the float is determined by Eq. (11.3).

When Table 11.3 is filled, it provides useful information about the entire set of jobs and may be a valuable aid for the manager in charge of the mission. This information in particular indicates for the case under consideration the following points.

a. The general mission time is 14 months. This is called the *lead time*.

b. Jobs E, F, G, and I have zero float and are referred to as *critical jobs*. Any delay in performing a critical job increases the mission time for the entire project. Therefore, maximum attention should be paid to these jobs.

c. Job D has a rather short float (1 month), therefore its progress should be painstakingly monitored.

d. Other jobs have considerable float and the manager may pay less attention to their progress.

To summarize, the table suggests that the manager should focus his attention on those jobs which are crucial for the entire project. However, tabular representation is inconvenient and less readily apparent than graphical representation.

11.2.4 Constructing a Network Diagram

In a graphical representation, called a network each job is shown as an arrow connecting two nodes signifying two extreme points of a job, as shown in Fig. 11.2*a*. An event is indicated by a circle with

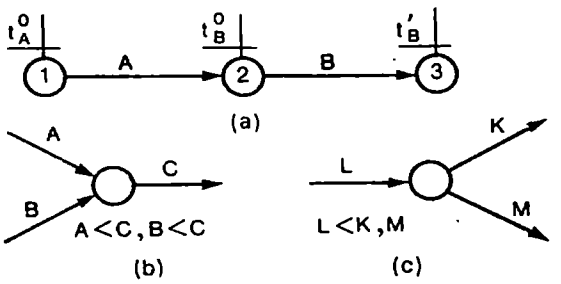


Fig. 11.2. (*a*) Graphical representation of jobs and events. (*b*) Union of jobs. (*c*) Splitting of jobs.

the instant of its occurrence being shown next to this node. Jobs may be labeled by letters or figures, the latter method being more convenient for computer analysis.

When a few jobs end with the same event, we say that a union of jobs takes place. If one event is a point of departure for a few jobs, then we say a job branching occurs. Examples of both situations are presented in Fig. 11.2.

The sequence of jobs in the network should reflect the constraints listed in Table 11.3. Let, for example, $L < K$, $M < K$, i.e., $\omega_k' = \{L, M\}$ and $t_M' < t_L'$. Then job K must begin at event I representing the end of job L with a latter termination time t_L', as shown in Fig. 11.3*a*. Job M ends ahead of L, therefore its end is associated with event 2. To reflect the condition $M < L$, events 2 and I are connected by broken-line arrow, implying *fictitious job* (2, 1) not

performed actually and representing just a float. This arguing brings about a number of dummy jobs in the network.

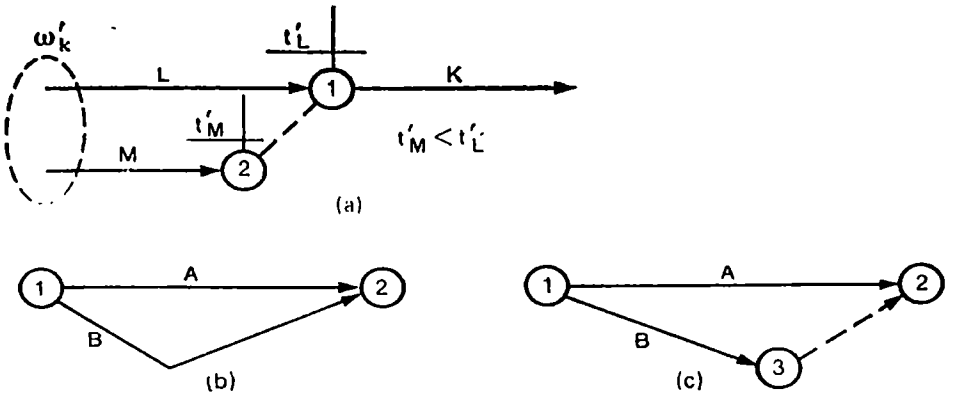
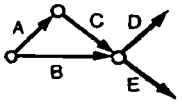
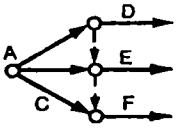
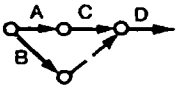
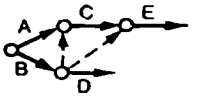
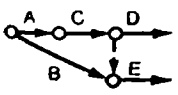
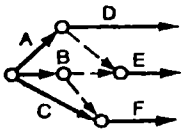
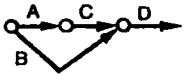
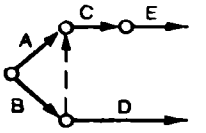


Fig. 11.3. (*a*) Introducing a dummy job implying an intention of float, also to discern between jobs (*b*) and (*c*)

Dummy jobs may not only represent floats, but also serve as an aid in separating some jobs indiscernible in numerical notation. For

Table 11.4

Redundant constraints		Fictitious jobs	
$A < C$	$A < D, E$	$A < C$	$A < C$
$B < E$	$B < E, F$	$B < D$	$B < C, D, E$
$C < D, E$	$C < F$	$C < D$	$C < E$
Incorrect			
			
Redundant constraint		No fictitious job needed	$B < E$ included twice
$B < D$	$A < F$		
Correct			
			

example, with reference to Fig. 11.3*b*, jobs *A* and *B* begin and end with the same events *1* and *2* and are indistinct in numerical notation thus preventing a computer analysis. In this case a dummy node *3* should be introduced along with dummy job (*3*, *2*) having a zero time span (Fig. 11.3*c*).

One should be cautious in introducing dummy jobs and events. Dummy jobs should be checked to see that they do not violate the

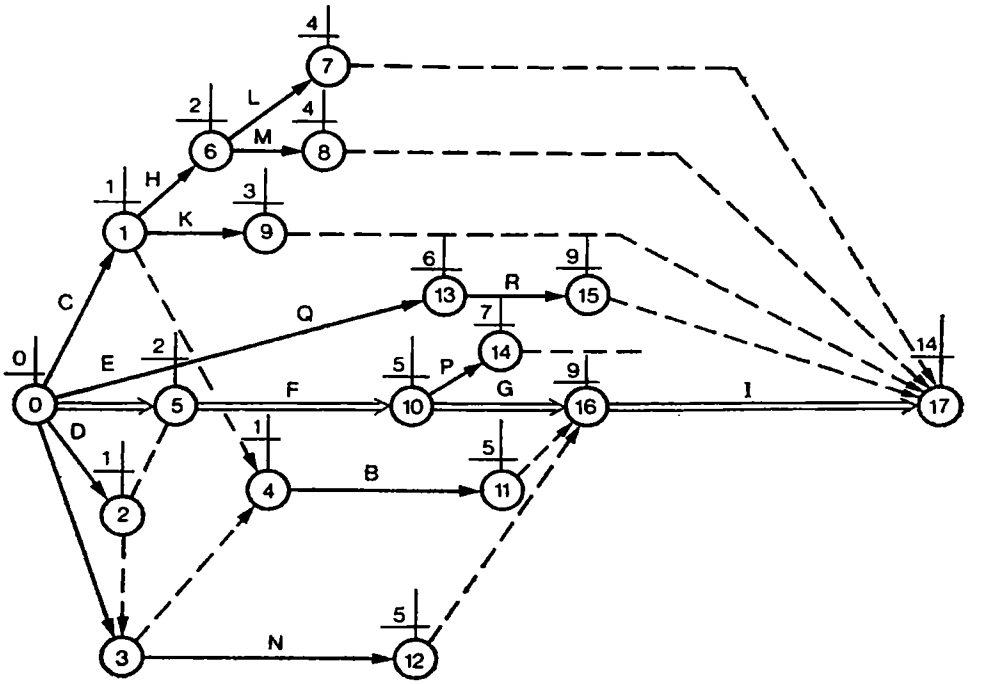


Fig. 11.4. Initial network

constraints imposed on the sequence order of jobs and do not introduce new constraints absent in the original table. Also no unnecessary and surplus dummy jobs should occur as these would introduce unnecessary complications in the analysis. Typical examples of incorrect and correct insertion of dummy jobs are presented in Table 11.4.

The network constructed by the data of Table 11.3 so that each job k begins at the earliest start time t_k^0 is presented in Fig. 11.4. The double-shafted arrows form a critical path that determines the duration of the entire project.

A network plan constructed by the outlined rules may not be totally satisfactory. It may offer an unsatisfactory allocation of resources to steps or too long a mission time or some other pitfalls. These deficiencies may be eliminated by optimization of the network. Some of the relevant procedures will be considered below.

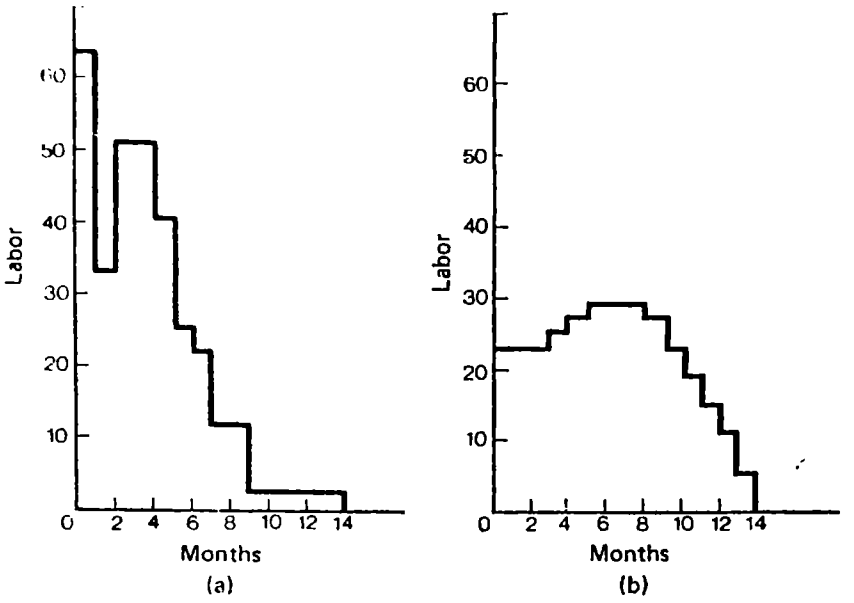


Fig. 11.5. Resource distribution

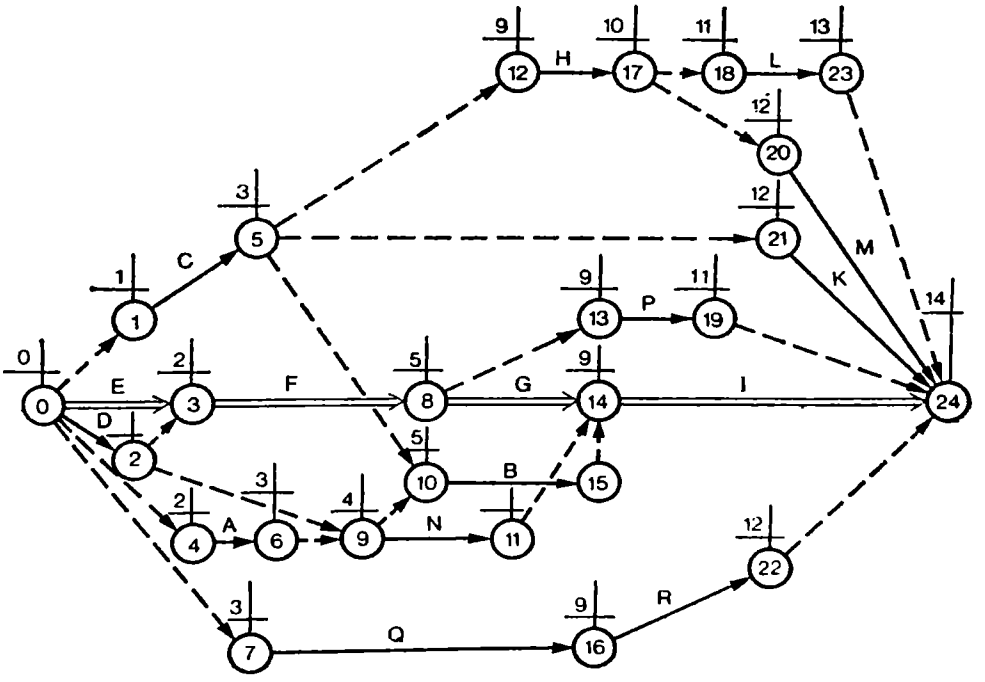


Fig. 11.6. Corrected network plan

11.2.5 Correction of Resource Allocation

If all jobs begin in the earliest possible time instants, the distribution of resources (labor requirement) depicted in Fig. 11.5*a* is quite nonuniform showing a figure of 62 in the first month and 32 in the second.

This distribution may be made more uniform by shifting the onsets of some jobs having considerable floats. A correction of the plan results from this shifting procedure. Figure 11.6 shows the corrected network corresponding to the distribution of resources presented in Fig. 11.5*b*.

11.2.6 Mission Time Optimization

It may so happen that the scheduled mission time T is too long and ways are sought to reduce this duration to $T' < T$. This goal may be achieved by reducing the duration of some constituting jobs.

Let τ_i represent the initially planned duration of job i. Investing additional funds x_i can change the duration of this job. If the funds are not very large and bounded, say by c_i, then the new duration of the job, τ_i', may be deemed linearly dependent upon x_i, i.e.

$$\tau_i' = \tau_i (1 - b_i x_i) \tag{11.9}$$

where b_i are coefficients known for each job.

A new formulation of the problem seeks to find additional investments x_i in each job such that reduce mission time to T'. The allocation of these funds should be such that the required reduction of terms is achieved at minimal investments. This problem is solved by the methods of linear programming.

Denote by v_i and t_i the start and finish times of job i. For the considered set of jobs, the constraints on order sequence give the following set of inequalities

$$v_i \geqslant 0, \; i, \; = A, \; \ldots, \; R \tag{11.10}$$

$$v_F \geqslant v_D, \; v_E; \; v_H \geqslant v_C, \text{ etc.} \tag{11.11}$$

The dependence of job duration on investments yields

$$t_i = v_i + \tau_i (1 - b_i x_i), \; i = A, \; \ldots, \; R \tag{11.12}$$

Constraints on the investments are as follows:

$$0 \leqslant x_i \leqslant c_i \; i = \mathrm{A}, \; \ldots, \; R \tag{11.13}$$

The condition that all the project should be completed by the given date implies that none of the end jobs should have an end date beyond T'', i.e.

$$t_i \leqslant T'', \; i = I, \; K, \; L, \; M, \; P, \; R \tag{11.14}$$

Now, we seek, subject to all these constraints, an allocation of additional investments x_i to jobs, such that the total sum of investments

$$L = \sum_i x_i \tag{11.15}$$

be a minimum.

This is a typical LP problem in which the unknowns are additional investments x_i and starting times for each job v_i. When the number of unknowns and constraints is large, the problem becomes extremely unwieldy. It may be simplified, however, if we observe that the total mission time is controlled by the duration of jobs constituting the critical path. If there are grounds to assume that the critical path remains the same for the new mission time, then the consideration should be carried out for jobs E, F, G, and I lying on this path.

11.2.7 Probability Consideration

The outlined methods of schedule graph construction rely on the assumption that the duration τ_i of each job is exactly known beforehand. In problems of actual value the durations of jobs depend upon many factors and are in fact random variables. Failure to observe this circumstance may lead to considerable errors in the evaluation of the mission time.

Estimation of time for each job is carried out by experts well versed in all the subtleties of the technology and organization of a particular job. When probabilistic methods are brought into the picture, the expert is required to yield three types of estimate

an optimistic estimate a_i giving the job time under most favorable conditions,

a pessimistic estimate b_i giving the job time under least favorable conditions.

a more likely time of job, m_i, that may take place under the existing conditions.

It is assumed that the job time is a random variable obeying a β-distribution law [69].

Now, the expected time of job is determined as

$$\bar{\tau}_i = (a_i + 4m_i + b_i)/6 \tag{11.16}$$

and the variance is

$$\sigma_i^2 = [(b_i - a_i)/6]^2 \tag{11.17}$$

All the other calculations are carried out as before but with the expected time of job in place of the specified time. The expected earliest and latest dates and floats for each job are determined. The difference from the earlier procedure consists in the simultaneous evaluation of variances of these times. The variance of a sum is calculated

as the sum of variances of the addends. For example, if the expected time of completion of job i is $\bar{t}_i = \bar{t}_j + \bar{\tau}_i$, then the variance of this term is $\sigma_i^2 = \sigma_j^2 + \sigma_\tau^2$.

Variance is a useful tool to estimate the probability of a job or the entire mission being completed on schedule. Suppose that the lead time is T_0. A calculation indicates that the expected mission time is $\bar{T}$ with variance σ_T^2. It would be reasonable to assume that the actual time of completion T will be a random variable with a normal distribution of mean $\bar{T}$ and variance σ_T^2, as shown in Fig. 11.7.

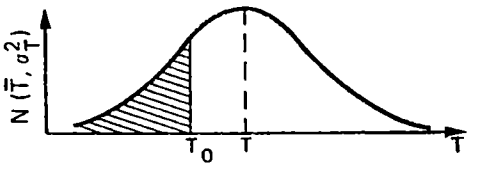


Fig. 11.7. Probability to complete a project on schedule by date T_0

The probability that the actual mission time T will not exceed the scheduled lead time T_0 is given by the shadowed area under the density curve. In agreement with (5.44) this probability may be found as

$$P(T \leqslant T_0) = \Phi[(T_0 - \bar{T})/\sigma_T] \qquad (11.18)$$

where $\Phi(x)$ is the standard normal cumulative distribution function tabulated as "normal integral".

The reader interested in a more detailed discussion of the aforementioned methods is referred to Refs. [31, 69, 75, and 76].

11.3 Combinatorial Scheduling Problems

11.3.1 Background

We illustrate the essence of combinatorial problems by an example of a job-shop process seeking an optimal sequence for parts processed on machines.

Assume that we have two machines, a lathe and a milling tool, and n parts each of which must be machined in sequence on each of these machines. We need to determine a processing sequence that allows a minimum idle time of the machines.

This problem reveals typical features of scheduling problems with constraints α on the sequence of operation (first turning, then milling operation) and constraints β on the available resources (one lathe and one milling machine). However, these constraints are not very strict and allow freedom of choice in choosing various processing sequences. Machining may be started with any of n parts, then any of the remaining $n - 1$ works can be processed and so on. The total number of versions is $n(n-1) \ldots 1 = n!$

Choosing various sequences we shall soon reveal that idle time of the machines depends on how the sequence is arranged, and waiting

time of works is also a function of the sequence. Striving to minimize idle time and waiting time we are obviously solving an optimization problem. However, its solution is rather difficult due to an enormous number of machining versions, relating this problem to the combinatorial class. We deter its solution until later. At this moment we consider a problem with simpler constraints that will introduce us to some solution techniques for this type of problem.

11.3.2 Traveling Salesman Problem

Consider a simpler version of the job-shop problem assuming that n different works numbered from 1 to n must be machined on the same tool. In coming from work i to work k a time d_{ik} is required for setup or changeover. Setup times are different for various i and k; they are summarized in Table 11.5. It is required to find a proc-

Table 11.5

i	k			
	1	2	...	n
1	×	d_{12}	...	d_{1n}
2	d_{21}	×	...	d_{2n}
.	.	.		.
.	.	.	...	.
.	.	.		.
n	d_{n1}	d_{n2}	...	×

essing sequence such that minimizes the overall processing time, also called flow time. Recognizing that for a work to be processed first there are n possibilities of choice, and there are $n - 1$ options for a second work the total number of combinations runs into $n!$ Consequently, this process relates to combinatorial problems.

Because the time spent for processing proper remains the same for any order of machining, the overall time minimization reduces to the minimization of the overall setup time.

In mathematics this type of problem is known as the *traveling salesman problem* (TSP) which in turn is a member of a more general class of mathematical programming problems. It can be simply stated as follows. Find a route for a traveling salesman who has to visit $(n - 1)$ cities in sequence and return to his home city, such that the total mileage of his journey is minimized and each city is visited exactly once. The distances between the cities d_{ik} are given in Table 11.5.

If we somewhat modify the initial formulation and shall not require that the traveling salesman should return to his home city, we obtain an *open* version of this problem. If we further assume that the home city is not specified, then we arrive at a complete analog to our job-shop problem. Indeed, each city corresponds to a job; the distance between cities corresponds to the changeover time for each pair of jobs, and the total travel distance corresponds to the total changeover time. These latter versions may be readily reduced to the original problem as will be demonstrated later.

11.3.3 Hamiltonian Path

The traveling salesman problem allows a fairly appealing representation in the form of a graph $G = (X, U)$ whose vertices correspond to cities, and edges to ways connecting these cities. A number d_{ik} called the *length of edge* i, k and equal to the distance between cities i and k is assigned to every edge i, k. If each city has a direct path to any other city, then the cities will be connected by edges in the graph representation. Vertices connected by edges are called adjacent. A graph in which every pair of vertices is adjacent is called a *complete graph.*

In practical problems complete graphs are rarely met. If a graph is not complete, it may be deemed one by mentally completing it with lacking edges of length $d = \infty$. In the distance table the respective cells remain void.

A simple path in graph G which passes through all vertices is known as a *Hamiltonian path.* A graph is said to be Hamiltonian if there is a simple circuit (called a *Hamiltonian circuit*) which passes through all the vertices of G. These names are associated with a game invented by the Irish mathematician Sir William Rowan Hamilton in the nineteenth century. It should be evident that the traveling salesman problem reduces to the evaluation of the shortest Hamiltonian circuit or path.

In what follows we shall predominantly deal with the closed version of the traveling salesman problem, i.e. one with the home city where the route starts and finishes. Open versions, however, may be readily reduced to closed.

Suppose that we have a graph $G = (X, U)$ and an initial point $x_0 \in X$, but the final point is not given. Replace the lengths of all edges coming into x_0 with zeros. Then any Hamiltonian circuit will contain a zero-length edge incident into x_0. If we find a shortest circuit and exclude from it a zero-length edge incident into x_0, then we obtain a shortest Hamiltonian path.

If neither initial nor terminal vertex of the graph is specified, then we introduce a dummy vertex z which is connected by zero-length edges with all vertices of the graph in both directions. We assume that

this vertex is the initial one and find a shortest Hamiltonian circuit. Eliminating from it the zero-length edge incident out of z and a zero-length edge incident into z, we obtain a shortest Hamiltonian path.

11.4 Examples of Problems Reducible to the Traveling Salesman Problem

11.4.1 The Hamilton Problem

Figure 11.8 shows a regular dodecahedron having twelve regular pentagons as faces and twenty vertices. The problem is to pass a string from one vertex to another (all marked by nails) passing through each vertex only once, covering all vertices, and returning to the initial vertex. Such a path, if existent, is called a *Hamiltonian circuit* or *cycle*.

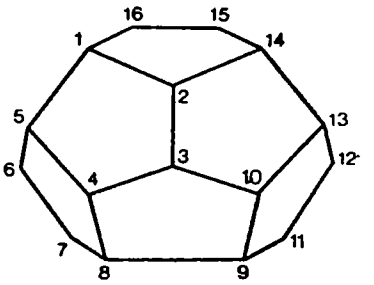


Fig. 11.8. Dodecahedron

To reduce this problem to the TSP we assume that the distance d_{ij} between two vertices equals unity if the vertices are adjacent, i.e. connected by an edge, and equals infinity otherwise. This reasoning may be used to compile a table of distances like Table 11.5 and solve the resulting TSP.

11.4.2 Transportation Problems

There is a large number of problems involving transportation from one point covering a number of other locations and return to the initial place. Their analogs are connections of a number of locations by power supply and other utility lines.

11.4.3 Software Optimization

Complex computations call for various operations performed by subroutines $B_1, \ldots, B_n$. As a rule a special program is written stating a certain order of addressing to these subroutines. However, in specific problems the order of these operations may be quite different. In such situations programming is reduced to introducing conditional branch instructions that violate the natural order of the program, so that operation B_i will be followed not by B_{i+1} but by some other operation B_k, as shown in Fig. 11.9.

Every branch instruction (also known as conditional transfer or conditional jump) takes additional time and complicates both programming and computing. Therefore, it would be desirable to write

the preliminary program so that handling a particular set of problems requires only a minimum number of conditional jumps.

Given a set of N problems to be solved on a computer in specific conditions. Denote by p_k the probability (frequency) of solving problem k. Define

$$\begin{aligned} a_{ij}^{(k)} &= 1, \text{ if } B_i \text{ is followed by } B_j \text{ in problem } k \\ &= 0, \text{ otherwise} \end{aligned} \tag{11.19}$$

Then the quantity

$$d_{ij} = \sum_{k=1}^{N} p_k a_{ij}^{(k)}; \quad i = 1, \ldots, n \tag{11.20}$$

is the expectation of the fact that under the circumstances operation B_i is followed by B_j. Conditionally d_{ij} may be taken to be the distance between B_i and B_j and operations proper may be treated as vertices of a graph. Then minimization of the average number of conditional jumps reduces to the evaluation of a shortest Hamiltonian path in the graph. The sequence of operation in this path suggests how the preliminary programming should be done.

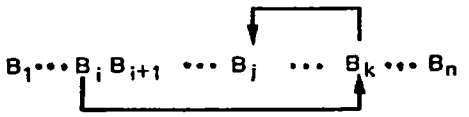


Fig. 11.9. Illustrating conditional branch instructions

11.4.4 Two-Machine Job-Shop Process

Here we revisit the problem of processing workpieces on two machines. Define a_k and b_k to be the processing time for work k on the first (A) and second (B) machines, respectively. We wish to determine a sequence of processing such that minimizes the total processing time, also called *flow time* or *makespan*.

To begin with we note the specific features of processing for each machine. The first machine, A, can begin processing of a next work immediately after finishing the predecessor work, so that it will have no idle time. The second machine, B, can begin processing of a successor work only when (1) the processing of this work by machine A has been completed, (2) the processing of a predecessor work has been completed by machine B.

Consider a simple situation involving two workpieces. If the first work is processed first, then the finish time of its processing by ma-

chine A will be a_1. The time points $a_1 + b_1$ and $a_1 + b_2$ correspond to the completion of processing the first work on machine B and the second work on machine A. The second work becomes available for processing when both these operations are completed, i.e. at the time $\max\ (a_1 + b_1, a_1 + a_2) = a_1 + \max\ (a_2, b_1)$. Processing of both workpieces completes at the time

$$t' = a_1 + \max\ (a_2, b_1) + b_2 \tag{11.21}$$

Likewise if the second work is processed first, then the instant of completion of both works is

$$t'' = a_2 + \max\ (a_1, b_2) + b_1 \tag{11.22}$$

Clearly, the processing should begin with the first work if $t' \ll t''$, i. e. if

$$a_1 + \max\ (a_2, b_1) + b_2 \leqslant a_2 + \max\ (a_1, b_2) + b_1 \tag{11.23}$$

If we recognize the obvious relations

$$a_1 + b_2 = \max\ (a_1, b_2) + \min\ (a_1, b_2)$$
$$a_2 + b_1 = \max\ (a_2, b_1) + \min\ (a_2, b_1)$$

then (11.23) may be rewritten as

$$\min\ (a_1, b_2) \leqslant \min\ (a_2, b_1) \tag{11.24}$$

This condition suggests when processing should begin with the first workpiece.

Consideration of a case with n workpieces is more involved, but it may be demonstrated that work i should be processed before work j if

$$\min\ (a_i, b_j) \leqslant \min\ (a_j, b_i) \tag{11.25}$$

If we put in correspondence to each work a vertex of a graph and define d_{ij}, the distance between vertices i and j, equal to unity when the condition (11.25) is satisfied and equal to infinity otherwise, then we arrive at a distance matrix specifying the respective traveling salesman problem. The shortest Hamiltonian path defined by this matrix yields the optimal sequence of processing.

11.5 Solution of Traveling Salesman Problems

11.5.1 Monte Carlo Simulation

Any statistical procedure involving random sampling from a probability distribution is called a Monte Carlo simulation model. We consider application of this method to solving traveling salesman problems.

Let vertex 1 be the initial vertex. Put in an urn balls labeled from 2 to n. After shaking the urn with the balls, balls are taken out one by one and their labels written into a sequence like $i_2, \ldots, i_n$. This procedure gives a Hamiltonian circuit $1, i_2, \ldots, i_n, 1$. The length of this circuit is computed and stored. Then the procedure is repeated and stored in place of the former if the new circuit is better, otherwise the new circuit is discarded. After multiple repetition, the analyst may expect obtaining with a high probability if not the best route, still a sufficiently good one.

Monte Carlo simulations are usually performed on computers equipped with random number generators, allowing one to look through a considerable number of routes in a short time.

11.5.2 Reduction to an Integer Program

Given a graph $G = (X, U)$ corresponding to a TSP, consider a Hamiltonian circuit $1, i_2, \ldots, i_n, 1$. The set of arcs constituting the circuit may be represented in matrix form

$$\mathbf{Y} = [y_{ij}] = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nn} \end{pmatrix} \tag{11.26}$$

where

$$\begin{aligned} y_{ij} &= 1, \quad \text{if arc } (i, j) \text{ is in the circuit} \\ &= 0, \text{ otherwise} \end{aligned} \tag{11.27}$$

Because each vertex occurs in a Hamiltonian circuit only once, then it has one and only one arc incident into it and one arc incident out of it. This implies that each row and each column of matrix (11.26) must have only one unity, which fact can be written as

$$\sum_{i=1}^{n} y_{ij} = 1, \quad \sum_{j=1}^{n} y_{ij} = 1 \tag{11.28}$$

In view of (11.28) the condition of all y_{ij} being 0s or 1s may be replaced with the requirement that all y_{ij} be nonnegative integers, namely,

$$y_{ij} \geqslant 0, \; i, j = 1, \ldots, n \tag{11.29}$$

$$\text{ent } [y_{ij}] = y_{ij} \tag{11.30}$$

where ent $[y_{ij}]$ means the integer part of the embracketed y_{ij}.

Now, any Hamiltonian circuit can be described by matrix (11.26) whose elements satisfy the relations (11.28-30). The length of a Hamiltonian circuit is equal to the sum of the constituting arcs

$$L = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij} \tag{11.31}$$

Thus, a Hamiltonian circle of minimal length can be evaluated by finding y_{ij} $(j = 1, \ldots, n)$ such that minimize the linear form (11.31) subject to (11.28) and (11.29). This is an integer programming problem.

It should be recognized, however, that the conditions (11.28-30), stating that each vertex has only one arc terminating at it and one arc leaving it, need not necessarily define a Hamiltonian circuit. These conditions will also be satisfied by a circuit that does not pass through all vertices if the remaining vertices have loops. This corresponds to some elements in the principal diagonal of matrix (11.26) being 1s and can be precluded by imposing additional constraints of the form

$$y_{ij} = 0 \text{ for } i = j,\ j = 1, \ldots, n \tag{11.32}$$

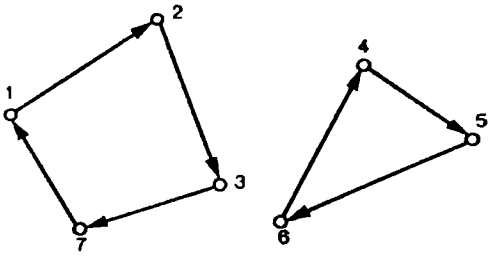


Fig. 11.10. Two non-connected circuits

Situations are also possible with two or more closed circuits, each of which embracing only a fraction of vertices, as shown for example in Fig. 11.10 for $n = 7$. If such a solution is anticipated, an additional constraint is introduced in the form

$$y_{12} + y_{23} + y_{37} + y_{71} \leqslant 3$$

for the considered situation. This condition prevents the occurrence of a circuit $(1, 2, 3, 7, 1)$ and the problem is solved a new subject to the new constraint.

11.5.3 Branch and Bound Algorithm

Let $U = \{1, \ldots, m\}$ be a set of solutions to some problem, labeled from 1 to m. Define q_i to be the value of a measure characterizing the performance of solution in version i. The optimal solution will be the one giving the minimal value

$$q^* = \min_{i \in U} q_i \tag{11.33}$$

to the performance measure, which is the exact lower bound of the set $\{q_1, \ldots, q_m\}$.

We define the *estimate from below* for a set U to be a value $q = \hat{q}$ satisfying

$$\hat{q} < q^* \quad \text{or} \quad \hat{q} \leqslant q^* \tag{11.34}$$

This estimate is said to be *attainable* if (11.34) is not a strict inequality, i.e. connected by the sign $\leqslant$.

In the framework of this definition the branch and bound method consists in a sequential improvement of $\hat{q}$ and approaching it to q^*. We assume that we can obtain an estimate from below for the entire set U and for its various subsets.

Let us partition U into two subsets, A and B, having the exact lower bounds q_A^* and q_B^* of the performance index related to q^* by

$$q^* = \min (q_A^*, q_B^*) \tag{11.35}$$

Because the sets A and B have fewer elements than U, there is a possibility of obtaining estimates $\hat{q}_A$ and $\hat{q}_B$ closer to q_A^* and q_B^* than in the initial set U. In other words, the estimate $\min (\hat{q}_A, \hat{q}_B)$ will be closer to q^* than $\hat{q}$.

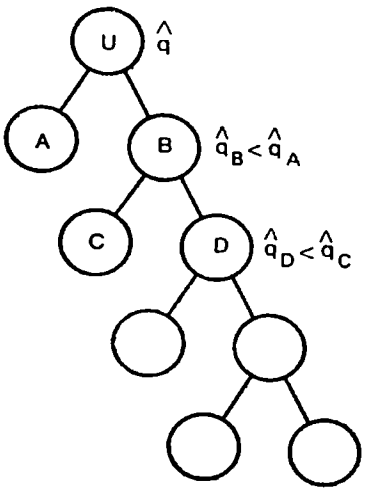


Fig. 11.11. Constructing a solution tree

There is also a high probability that the optimal version, i.e. the one with the smaller value of $q = q^*$, appears in a subset having the smaller estimate from below. This probability will be higher the farther from one another estimates $\hat{q}_A$ and $\hat{q}_B$ lie. Therefore, the set U should be partitioned into A and B so that their lower estimates differ as far as possible.

If it turned out that $\hat{q}_B < \hat{q}_A$, then there are grounds to believe that the optimal version belongs to B and this subset should be investigated in more detail. To this end, this set is partitioned into two parts C and D such that the respective lower estimate $\hat{q}_C$ and $\hat{q}_D$ differ as far as possible. If $\hat{q}_D < \hat{q}_C$, then we continue the partition procedure with the set D, and so on, until we arrive at a subset consisting of a single element with the value of the performance index $q = q_0$. For technical convenience the partition procedure is represented by a solution tree, as shown in Fig. 11.11.

It should be recognized, however, that the element with $q = q_0$ may only be expected to be optimal. It is required to check and see whether the rejected sets A, C, etc. have an element with $q < q_0$. Of course, sets with $\hat{q} > q_0$ may not be taken into consideration because they have no element with $q < q_0$.

Separation of the aforementioned sets either will lead to an element with $q < q_0$, then this element should be taken as an optimal,

or will lead to estimates from below $\hat{q} > q_0$. In this case the element with the objective value q_0 will be the desired solution of the problem.

11.5.4 Solving TSPs by the Branch and Bound Method

Suppose that a traveling salesman problem (TSP) is given by the distance matrix, like that of Table 11.5. To get specific, we shall illustrate our consideration with the matrix of Table 11.6a where the

Table 11.6

(a)

i \ j	1	2	3	4	5	
1	×	7/5	7/4	2/0	4/2	2/—
2	6/5	×	2/0	4/3	1/0	1/—
3	6/5	1/0	×	9/8	2/1	1/—
4	2/0	3/1	8/5	×	5/3	2/—
5	6/5	1/0	2/0	5/4	×	1/—

(b)

i \ j	2	3	4	5
1	5/3	4/2	×	2/0
2	×	0/0	3/0	0/0
3	0/0	×	8/5	1/1
5	0/0	0/0	4/1	×

(c)

i \ j	2	3	4
2	×	0	0
3	0	×	5
5	0	0	×

denominator of fractional quantities represents values for the *reduced* table.

To arrive at an estimate from below for the length of all Hamiltonian circuits we make use of the fact that each Hamiltonian circuit has only one element from each row and column of the distance matrix. Therefore, if we reduce the elements of a row or column by a number, then the lengths of all Hamiltonian circuits diminish by the same number. The reduction technique for the distance matrix relies upon this property.

The matrix can be reduced by rows and by columns. In reduction by rows, the least element of row i, denoted by h_i, is subtracted from all elements of the row. Clearly, the lengths of all Hamiltonian circuits diminish by the sum of values subtracted from each row, $\Sigma_i h_i$, while remaining nonnegative values. Therefore, the quantity $\Sigma_i h_i$, equal to 7 in our example, gives some estimate from below for the lengths of all Hamiltonian circuits. This estimate can be improved by reducing the distance matrix by columns. The least element g_j of column j is subtracted from the elements of this column. The sum of subtracted elements is $\Sigma_j g_j = 1$. The improved estimate from below of the lengths of Hamiltonian circuits will be

$$\hat{q} = \sum_i h_i + \sum_j g_j = 7 + 1 = 8$$

The reduced distance matrix is summarized (as denominators) in Table 11.6a.

This method of matrix reduction will be used below to obtain lower estimates for various subsets of the set of Hamiltonian circuits.

Consider now the method of partitioning the set of Hamiltonian circuits into subsets. Take an arc (i, j). Assign to the first subset all Hamiltonian circuits containing this arc. Denote this set by $[(i,j)]$. Assign to the second subset all Hamiltonian circuits which do not contain this arc. Denote this set by $[(i, j)]^C$.

The distance table for the set $[(i, j)]$ can be readily obtained by observing that if arc (i, j) is included in a Hamiltonian circuit, then this precludes inclusion of other arcs from row i or column j. Consequently, the distance table for $[(i, j)]$ results from the original table by deleting row i and column j.

In order to obtain the distance table for the set $[(i, j)]^C$ we need to block traffic through arc (i, j) in the original table by letting the length of this arc equal to infinity. This blocking of traffic through arc (i, j) will be denoted by a cross in the respective cell of the initial table.

A natural question arises: which game should be used as a basis for partitioning the set of routes into subsets? Note, above all, that the sets $[(i, j)]$ and $[(i, j)]^C$ are not identical in the number of elements. The total number of Hamiltonian circuits in the problem with n cities is $(n - 1)!$. If we take an arc (i, j), i.e. having selected two vertices, then there are $n - 2$ options to select a third vertex, $n - 3$ options to select a fourth, and so on. Consequently, there are $(n - 2)!$ Hamiltonian circuits including arc (i, j), and hence $(n - 1)! - (n - 2)! = (n - 2)(n - 2)!$ Hamiltonian circuits not including arc (i, j). Therefore, for $n > 2$, the set $[(i, j)]^C$ will contain more elements than the set $[(i, j)]$. Recognizing that a set with lesser number of elements is simpler to investigate, we observe that it would be wise to take as (i, j) an arc for which the set $[(i, j)]$ has a smaller estimate than the set $[(i, j)]^C$. Among the arcs satisfying this condition, we should prefer the one giving a greater difference in the estimates for $[(i, j)]$ and $[(i, j)]^C$.

If we take an arc (i, j) for which $d_{i,j} \neq 0$ in the reduced table, then for the set $[(i, j)]$ the estimate from below increases by $d_{i,j}$ because we know beforehand that this arc will enter all Hamiltonian circuits. It can increase still more if after deletion of row i and column j the table will allow a further reduction. At the same time, if we replace element $d_{i,j}$ with infinity, then the distance table remains reduced, i.e. the estimate from below for $[(i, j)]^C$ does not increase. Hence, a smaller estimate from below will be for the set $[(i, j)]^C$ which is undesirable. Therefore, arc (i, j) should be selected such that has $d_{i,j} = 0$ in the reduced table.

There are a few such arcs—which of them should be taken? Evidently the one for which an increase of the estimate for the set $[(i, j)]^C$ will be greatest, as this results in a largest difference in the estimates for $[(i, j)]$ and $[(i, j)]^C$.

Denote the increase of the estimate for $[(i, j)]^C$ by $\Delta(i, j)$. This quantity is obtained as the sum of the least number in the ith row and jth column.

Let us return to our example; for Table 11.6a (denominator), the values of $\Delta(i, j)$ are as follows:

$$\Delta(1, 4) = 2 + 3 = 5 \qquad \Delta(3, 2) = 1$$
$$\Delta(2, 3) = 0 \qquad \Delta(4, 1) = 6$$
$$\Delta(2, 5) = 1 \qquad \Delta(5, 3) = \Delta(5, 2) = 0$$

The largest increase of the estimate for $[(i, j)]^C$ results for arc (4,1). Deleting the fourth row and first column from the reduced matrix, we obtain the matrix of Table 11.6b (numerator) representing the distance matrix for the set $[(4, 1)]$.

In handling such matrices one should carefully check to see that no circuits could occur that are not Hamiltonian. Since arc (4, 1) is already included in the selected Hamiltonian circuit, we should block the passage through arc (1, 4) and mark this cell by a cross, as arc (1,4) along with (4,1) forms a circuit. Table 11.6b (denominator) is the reduced distance matrix with an increase of estimate of $\hat{\Delta q} = 2 + 3 = 5$.

For Table 11.6b (denominator), the values of $\Delta(i, j)$ are as follows:

$$\Delta(1, 5) = 2 \qquad \Delta(2, 5) = 0$$
$$\Delta(2, 3) = 0 \qquad \Delta(3, 2) = 1$$
$$\Delta(2, 4) = 1 \qquad \Delta(5, 2) = \Delta(5, 3) = 0$$

The largest increase of the estimate for the set $[(i, j)]^C$ results for arc (1, 5) which will be included also into the sought Hamiltonian circuit, thus obtaining the path (4, 1, 5). Eliminating the first row and fifth column from the matrix of Table 11.6b (denominator), we obtain Table 11.6c for the set $[(4, 1)\ (1, 5)]$. In this table we should forbid arc (5, 4) making a circuit with the path $[(4, 1, 5)]$.

The resultant table is reduced, therefore we immediately find the values of $\Delta(i, j)$

$$\Delta(2, 3) = 0 \qquad \Delta(3, 2) = 5$$
$$\Delta(2, 4) = 5 \qquad \Delta(5, 2) = \Delta(5, 3) = 0$$

Thus, the largest values of $\Delta(i, j)$ are equal to arcs (2,4) and (3,2). Therefore, both arcs may be included in the sought Hamiltonian circle thus arriving at the arc sequence (3, 2), (2, 4), (4, 1), (1, 5) which lacks arc (5, 3) to be closed. Adding this arc, we obtain the assumed optimal Hamiltonian circle (3, 2, 4, 1, 5, 3) of length 13.

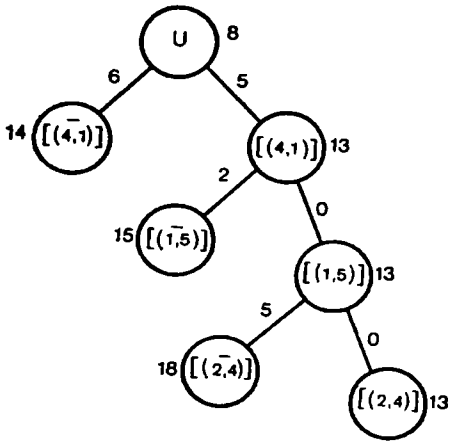


Fig. 11.12. Solution tree in the traveling salesman problem

Analysis of the subsets left out of consideration, $[(4, 1)]^C$, $[(4, 1), (1, 5)^C]$, $[(4, 1), (1, 5), (2, 4)^C]$ indicates that their estimates from below are higher than 13, therefore these subsets cannot contain a Hamiltonian circuit shorter than that found above. Consequently, further check is not necessary and the Hamiltonian circuit thus obtained is optimal.

Figure 11.12 shows the solution tree constructed for this problem.

More detailed discussions of solution techniques for combinatorial problems may be found in Refs. [73, 77, and 78].

11.6 Theory of Queues

11.6.1 Basic Concepts and Definitions

A "queue" is a waiting line of units demanding service at a service facility. Queueing problems arise in the systems where units, called

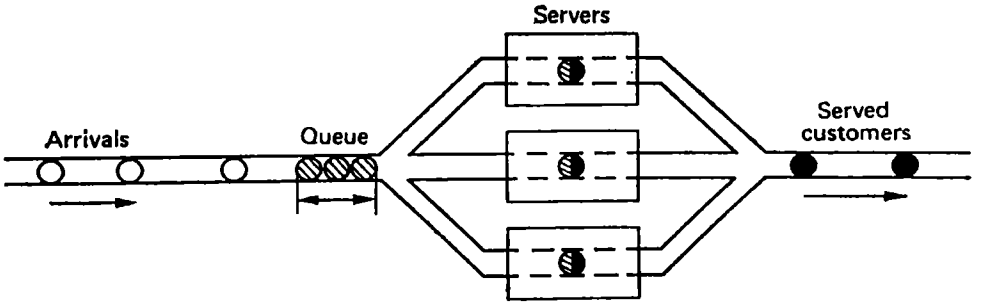


Fig. 11.13. Major elements of a queueing system

customers or *arrivals*, arrive for service at random time instants and get service from one or a few *servers* or *service channels*. The time during which service is rendered to individual customers is also assumed random. The classical model of a queue with three service channels is depicted in Fig. 11.13. In a multichannel system all channels or some of them may render one and the same or different types of service.

The rate at which customers arrive to be serviced is called the *arrival rate*, λ. The rate at which the service unit can provide customer service is called the *service rate*, μ. Typical situations with queueing systems are those occurring because of mismatch of the arrival rate and service rate and manifesting themselves in servers being idle or customers queueing up in line for service. Examples from everyday life are people in queues for tickets, cars waiting for gas and oil at gas stations, airplanes waiting for the runway.

The principal characteristics of queueing systems are as follows.

Number of customers in the queue at various time instants.

Expected system time. The expected time a customer spends waiting plus being serviced.

The probability of an empty or idle system.

To build a mathematical model of a queueing system we need (1) a specification of the *input process* including the source of customers, the type of customers, and the inter-arrival times; (2) a specification of the *service mechanism* including the number of servers, the number of customers getting served at any time, and the duration and mode of service; (3) a description of a *queue discipline* indicating how customers arrive from the queue at a service unit including such rules as "first-come, first-served", "last-come, first-served", or some priority disciplines.

In queueing system studies, the input process of customers is usually assumed to be distributed in a Poisson law (see Sect. 5.7) with arrival rate λ. This implies that customers arrive at random time instants so that the probability of an arrival in the interval from t to $t + \Delta t$ is equal to $\lambda \Delta t$ and is independent of t, while the probability of arrival of two and more customers in this interval is negligibly small. These assumptions are fairly justified for many practical problems.

We assume that service times are distributed exponentially with mean $1/\mu$. This means that the probability of a service being completed in the time interval from t to $t + \Delta t$ is independent of t and is equal to $\mu \Delta t$. This distribution does not always provide a good description of actual service unit performance, but its utilization is justified by a rather simple mathematical analysis. In what follows we also assume that the queue discipline is first-come, first-served.

11.6.2 Differential Equations of the Model

Denote by S_i the state of the queueing system with i customers in the queue ($i = 0, 1$, etc.) and by $p_i(t)$ the probability that the system is in state S_i at time t. Assume also that the arrival and service rates, λ_i and μ_i, depend in the general case on the number of customers in the queue.

If we take N identical queueing systems, then at time t a proportion $p_{i-1}(t)\,N$ of them will be in state S_{i-1}, $p_i(t)\,N$ in state S_i, and $p_{i+1}(t)\,N$ in state S_{i+1}. Let us investigate how the number of systems $\Delta p_i(t)\,N$ residing in state S_i varies in the time interval from t to $t+\Delta t$.

Given a Poisson input process and exponentially distributed service times, no more than one customer can arrive and get service within a small time interval Δt in either of the systems, i.e. the state of each system cannot change by more than one unit. Because the probability of arrival and service completion in Δt are $\lambda\Delta t$ and $\mu\Delta t$, respectively, then within this interval $\lambda_i\Delta t p_i(t)\,N$ systems in state S_i receive new customers and find themselves in state S_{i+1}, while $\mu_i\Delta t p_i(t)\,N$ systems complete servicing of a sequential unit and move to state S_{i-1}. In addition, $\lambda_{i-1}\Delta t p_{i-1}(t)\,N$ systems being in state S_{i-1} move to state S_0 after a next customer has arrived, and $\mu_{i+1}\Delta t p_{i+1}(t)\,N$ systems in state S_{i+1} having serviced a customer move to state S_i. The variation of the total number of systems $\Delta p_i(t)\,N$ in state S_i within Δt will be

$$\begin{aligned}\Delta p_i(t)\,N = &-\lambda_i\Delta t p_i(t)N-\mu_i\Delta t p_i(t)\,N\\ &+\lambda_{i-1}\Delta t p_{i-1}(t)\,N+\mu_{i+1}\Delta t p_{i+1}(t)\,N\end{aligned}$$

Dividing this equation by Δt and letting $\Delta t\to 0$, we have

$$\frac{dp_i(t)}{dt}=-(\lambda_i+\mu_i)\,p_i(t)+\lambda_{i-1}p_{i-1}(t)+\mu_{i+1}p_{i+1}(t) \qquad (11.36)$$

This equation is inapplicable at $i=0$, as in this case customers can not leave the system and $p_{i-1}(t)$ becomes senseless. Therefore this case requires a separate consideration. At $i=0$ the system can change its state S_0 for S_1 by receiving a new customer, and state S_1 for S_0 by completing a single customer, therefore,

$$dp_0(t)/dt=\lambda_0 p_0(t)+\mu_1 p_1(t) \qquad (11.37)$$

Equations (11.36-37) constitute the "birth and death" model used for population studies in biology. In queueing theory, given the initial conditions and rates λ_i and μ_i, these equations define the variation of the character of service in time. However, for practically important situations these equations are extremely difficult to solve.

As a rule the transient process just described is of less interest than the stationary state of the queueing system called the *statistical equilibrium*. In this state, probabilities $p_i(t)$ do not vary in time so that $dp_i(t)/dt=0$ and the system model is as follows:

$$\begin{aligned}\lambda_0 p_0&=\mu_1 p_1\\ (\lambda_i+\mu_i)\,p_i&=\lambda_{i-1}p_{i-1}+\mu_{i+1}p_{i+1}\end{aligned} \qquad (11.38)$$

Letting $i=1, 2$, etc. and subtracting each equation from the predecessor, we obtain the system of equations for statistical equi-

librium

$$\begin{aligned} \lambda_0 p_0 &= \mu_1 p_1 \\ \lambda_1 p_1 &= \mu_2 p_2 \\ &\cdots\cdots\cdots \\ \lambda_i p_i &= \mu_{i+1} p_{i+1} \end{aligned} \tag{11.39}$$

The solution to this system has the form

$$\begin{aligned} p_1 &= (\lambda_0/\mu_1)\, p_0 \\ p_2 &= (\lambda_0\lambda_1/\mu_1\mu_2)\, p_0 \\ &\cdots\cdots\cdots\cdots \\ p_i &= (\lambda_0\lambda_1,\ldots\lambda_{i-1}/\mu_1\mu_2\ldots\mu_i)\, p_0 \end{aligned} \tag{11.40}$$

It should be evident that this steady-state mode of operation can actually exist provided that

$$\sum_{i=0}^{\infty} p_i = 1 \tag{11.41}$$

Below we consider application of (11.40-41) to the description of some popular types of queueing systems.

11.6.3 Multichannel Queue with no Waiting Room

Consider a queueing system with n identical service channels and an input process of arrival rate λ. If an arrival finds a free channel in the system, it gets served, otherwise it leaves the system unserved. This situation occurs where customers cannot form a queue as in telephone exchange where a call cannot be passed through if all the channels are busy and there is no such instrument as a waiting list.

In these situations we shall understand the length of queue i as the number of busy channels so that the arrival rate is

$$\begin{aligned} \lambda_i &= \lambda, \quad \text{if } i \leqslant n \\ &= 0, \quad \text{if } i > n \end{aligned} \tag{11.42}$$

If μ is the service rate of one channel, then the service rate with i channels busy

$$\mu_i = i\mu, \qquad i = 1, \ldots, n \tag{11.43}$$

The probability that i channels are busy can be found by Eqs. (11.40) as

$$p_i = \lambda^i p_0/i!\mu^i = \rho^i p_0/i! \tag{11.44}$$

where the ratio of effective arrival and service rates

$$\rho = \lambda/\mu \tag{11.45}$$

is called *traffic intensity*.

The probability that all channels are idle, p_0, can be found by (11.41) as

$$p_0 = \left(1 + \rho + \frac{1}{2!}\rho^2 + \ldots + \frac{1}{n!}\rho^n\right)^{-1} \tag{11.46}$$

System operation is usually characterized by the probability of loss, p_{loss}, and the expected number of busy channels, $\bar{k}$. The probability of loss of a customer is equal to the probability of all channels being busy

$$p_{\text{loss}} = p_n = \rho^n p_0/n! \tag{11.47}$$

At times the concept of relative throughput of the system is more convenient to use than the probability of a busy system. It is defined as the probability that a call will be served

$$c = (1 - p_{\text{loss}})\lambda \tag{11.48}$$

The expected number of busy channels is

$$\bar{k} = \sum_{i=1}^{n} i p_i = p_0 \sum_{i=1}^{n} \frac{\rho^i}{(i-1)!} \tag{11.49}$$

Example 11.1. Consider a no-waiting-room multi-channel system with $\rho = \lambda/\mu = 2$.

For a three-channel queue

$$p_0 = (1 + \rho + \rho^2/2 + \rho^3/6)^{-1} = 0.16$$

The probability of a busy system (loss of customers) is

$$p_{\text{loss}} = p_3 = \rho^3 p_0/6 = 0.21$$

The expected number of busy channels is

$$\bar{k} = p_0 (\rho + \rho^2 + \rho^3/2) = 1.58$$

For a two-channel queue, $p_{\text{loss}} = p_2 = 0.4$; and $\bar{k} = 1.2$.

11.6.4 Single Channel Queue

A system handles customers arriving at a rate λ with service rate μ in a single channel. The balance of state equations for this queue can be obtained from (11.40) by letting $\lambda_i = \lambda$, $\mu_i = \mu$, and $\rho = \lambda/\mu$, namely,

$$p_i = \rho^i p_{0i} \qquad i = 1, 2, \text{ etc.} \tag{11.50}$$

For this case, Eq. (11.41) takes the form

$$p_0 (1 + \rho + \rho^2 + \rho^3 + \ldots) = 1$$

that is meaningful only for $\rho < 1$. Summing up the parenthesized geometric sequence, we get

$$p_0 = 1 - \rho \tag{11.51}$$

so that

$$p_i = (1 - \rho)\,\rho^i,\ i = 0, 1, 2, \text{etc.} \tag{11.52}$$

The principal characteristics of this queue are the number in the queue and the expected time in the queue. The expected number of customers in the system (queue and service) is

$$\bar{k} = (p_1 + 2p_2 + 2p_3 + \ldots) = (1 - \rho)\,(\rho + 2\rho^2 + 3\rho^3 + \ldots)$$

If we recognize that [70]

$$\rho + 2\rho^2 + 3\rho^3 + \ldots = \rho/(1 - \rho)^2$$

then

$$\bar{k} = \rho/(1 - \rho) \tag{11.53}$$

The expected number of customers being serviced can be easily found by noting that this is equal to 1 times the probability that the service unit is busy, or $1 \cdot \lambda/\mu$, viz.,

$$\bar{k}_s = \lambda/\mu = \rho \tag{11.54}$$

The difference between $\bar{k}$ and $\bar{k}_s$ gives the expected number in the queue

$$\bar{k}' = \bar{k} - \bar{k}_s = \rho/(1 - \rho) - \rho = \rho^2/(1 - \rho) \tag{11.55}$$

If we observe that the average service time of one arrival is $1/\mu$, then the expected time in the queue is

$$t_q = \bar{k}/\mu \tag{11.56}$$

11.6.5 Multiple-Channel Queue

If the queueing system consists of n channels handling an input process arriving at a rate λ with a service rate μ in each channel, then

$$\lambda_i = \lambda,\ i = 0, 1, 2, \text{etc.} \tag{11.57}$$

$$\mu_i = i\mu \text{ for } i \leqslant n = n\mu \text{ for } i > n \tag{11.58}$$

It will be demonstrated later that this system has a statistical equilibrium if $\rho = \lambda/\mu < n$. In this case

$$p_i = \lambda^i p_0/(\mu \cdot 2\mu \cdot \ldots \cdot i\mu) = (\rho^i/i!)\,p_0,\ i \leqslant n \tag{11.59}$$

$$p_i = [\lambda^i/\mu \cdot 2\mu \cdot \ldots \cdot n\mu \cdot (n\mu)^{i-n}]\,p_0$$

$$= (\lambda^i/n!\,\mu^n n^{i-n})\,p_0 = (\rho^i/n!\,n^{i-n})\,p_0 = \frac{\rho^n}{n!}\frac{\rho^{i-n}}{n^{i-n}}p_0,\quad i > n \tag{11.60}$$

For these values of p_i, Eq. (11.41) takes the form

$$p_0\left[1+\rho+\frac{\rho^2}{2!}+\ldots+\frac{\rho^{n-1}}{(n-1)!}+\frac{\rho^n}{n!}\left(1+\frac{\rho}{n}+\frac{\rho^2}{n^2}+\ldots\right)\right]=1$$

Recognizing that

$$1+\frac{\rho}{n}+\frac{\rho^2}{n^2}+\ldots=\frac{1}{1-\rho/n}=\frac{n}{n-\rho}$$

we get

$$p_0=1\Big/\left(1+\rho+\frac{\rho^2}{2!}+\ldots+\frac{\rho^{n-1}}{(n-1)!}+\frac{\rho^n}{(n-1)!\,(n-\rho)}\right) \tag{11.61}$$

The key parameter of the system may readily be found by employing the line of argument of the single channel model [74]

$$\bar{k}'=\frac{\rho^{n+1}}{nn!\,(1-\rho/n)^2}\,p_0=\frac{\rho^{n+1}}{(n-1)!\,(n-\rho)^2}\,p_0 \tag{11.62}$$

$$\bar{k}_s=\rho \tag{11.63}$$

$$\bar{k}=\bar{k}'+\rho \tag{11.64}$$

$$t_q=\bar{k}/\mu n \tag{11.65}$$

These models of infinite population queues cannot, of course, cover all the multitude of practical queueing systems and have been elucidated to demonstrate the general approach to analysis of such situations. Much more space is devoted to this topic in Refs. [31, 32, 71, 74, and 79].

11.6.6 Simulation of Queueing Systems

Analytical methods are applicable only for comparatively simple queueing systems. Therefore in analysis of actual cases with input process and service times obeying other distributions than those yielding the simple aforementioned models analysts have to resort to simulation of queues on a computer. The simulation process relies upon the generation of arrival times and service times by random number generators wired in computing systems.

Sequences of random numbers γ_n, $n = 1, 2, \ldots$, uniformly distributed in the interval from 0 to 1 are simpler to realize. As a rule, these numbers are obtained by the recurrence formulas

$$\gamma_n=\Phi\,(\gamma_{n-1}),\ n=1,\ 2,\ \text{etc.} \tag{11.66}$$

for a given initial value γ_0. We restrict our attention to one of such algorithms in which random numbers γ_n are obtained by

$$\gamma_n = g\gamma_{n-1} - \text{ent}\,(g\gamma_{n-1}),\ n = 1, 2, \text{etc.} \qquad (11.67)$$

where g is an integer far greater than unity, ent $(g\gamma_{n-1})$ is the integer part of $g\gamma_{n-1}$, and γ_0 is a given number.

In fact this algorithm produces so-called pseudo random numbers, and being realized on a computer it always yields periodic sequences. Indeed, because in a computer numbers are represented with a finite number of places, the total of numbers between zero and unity will be finite. This implies that sooner or later some value γ_L will coincide with one of its predecessors γ_i. Then by virtue of (11.67)

$$\gamma_{L+i} = \gamma_{l+i} \text{ for } i = 1, 2, \text{etc.} \qquad (11.68)$$

Let L be the smallest number satisfying (11.68). Then L is the aperiodic length and $p = L - l$ is the length of the period. It is quite natural that simulation should use not more than L numbers of the sequence (11.67). Therefore, once the numbers g and γ_0 have been chosen, the analysis should experimentally determine the aperiodic length, and check to see how close the distribution of the resulting pseudo random sequence is to the desired uniform distribution. Relevant verification techniques are described in Ref. [80].

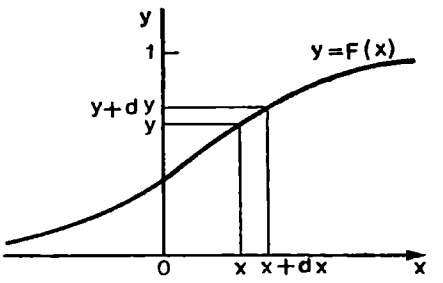


Fig. 11.14. Generation of a random sample with a given distribution

Simulation requires samples of random numbers from various distribution laws, which can be obtained by functional transformation of a sample from a uniform distribution. Consider some methods of solving this problem.

We wish to develop an algorithm generating sequences of random numbers $x_1, x_2, \ldots, x_n$, etc. with a given probability density $w(x)$ or cumulative distribution function $F(x)$ related to $w(x)$ by Eq. (5.37).

Define $F(x) = y$ and construct the plot $y = F(x)$ as shown in Fig. 11.14. If we vary x from $-\infty$ to ∞, y will vary from 0 to 1. Now, we assume that y is a random variable uniformly distributed over [0, 1], i.e. having the probability density $p(y) = 1$ for $0 \leqslant y \leqslant 1$, and seek for the probability of $F(x)$ lying in the interval from y to $y + dy$, viz.,

$$P(y \leqslant F(x) < y + dy) = p(y)\,dy = dy = dF(x) = w(x)\,dx \qquad (11.69)$$

If $y = F(x)$ is uniformly distributed over [0, 1], then x has the probability density $w(x) = dF(x)/dx$. Therefore, in simulating a random process x with a given distribution $F(x)$, the numbers y should be random numbers γ generated by a random number generator with a uniform probability distribution over [0, 1], i.e. letting $\gamma = F(x)$, whence

$$x = F^{-1}(\gamma) \tag{11.70}$$

where $F^{-1}(\gamma)$ is an inverse function.

This method of generation of a sample of x with a given distribution function $F(x)$ is referred to as the *inverse transformation method.*

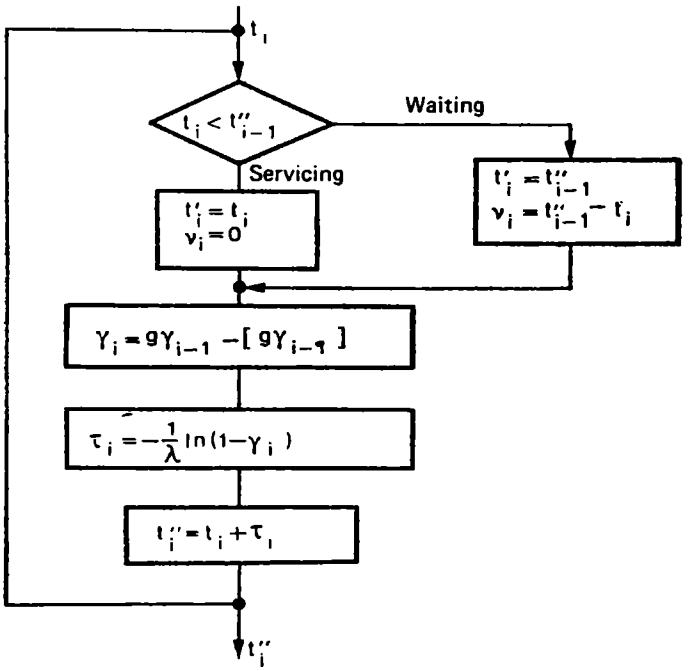


Fig. 11.15. Queue simulation algorithm

Example 11.2. Problems of reliability theory and analysis of queues make wide use of random variables x obeying an exponential distribution with density $w(x) = \lambda \exp(-\lambda x)$ and c.d.f.

$$F(x) = \int_{-\infty}^{x} w(\xi)\, d(\xi) = 1 - \exp(-\lambda x) \tag{11.71}$$

Assuming $F(x) = \gamma$, we find $\gamma = 1 - \exp(-\lambda x)$, whence

$$x = -(1/\lambda) \ln(1 - \gamma) \tag{11.72}$$

Since γ is distributed uniformly in the interval from 0 to 1, then $1 - \gamma$ will have the same distribution. Therefore, γ may be substituted for $1 - \gamma$ so that $x = -(1/\lambda) \ln \gamma$.

Example 11.3. A queueing system handles a regular input process with interarrival intervals of 2 minutes. The service time is exponential with rate $\lambda = 0.5$.

Simulation of system operation was performed on a computer by an algorithm displayed in Fig. 11.15. For an arrival i, the notation is as follows: t_i, arrival time; t_i', service starting time; v_i, waiting time; τ_i, time in service; and t_i'', service finishing time. A sample γ_i from a uniform distribution was obtained by Eq. (11.67) for $g = 3125$ and $\gamma_0 = 0.123456789$. The time in service, τ_i, was found by (11.72). Simulation data are set out in Table 11.7.

Table 11.7

i	t_i	v_i	t_i'	γ_i	τ_i	t_i''
1	2	0	2	0.8025	3.244	5.244
2	4	1.244	5.244	0.7037	2.433	7.677
3	6	1.668	7.677	0.2207	0.499	8.175
4	8	0.175	8.175	0.5542	1.615	9.791
5	10	0	10	0.9202	5.057	15.057
6	12	3.057	15.057	0.7572	2.831	17.889
7	14	3.889	17.889	0.3044	0.726	18.615
8	16	2.615	18.615	0.2424	0.555	19.170
9	18	1.710	19.170	0.6465	2.080	21.250
10	20	1.250	21.250	0.5021	1.395	22.645

Note: i = unit number; t_i = instant of arrival; v_i = waiting time; t_i' = start of service; γ_i = sample of uniform distribution; τ_i = service time; t_i'' = end of service.

The inverse function method is not the only possible in generating a random sample with a given distribution. Techniques have been developed [80, 81] that proved to be more convenient in many situations. We give relationships that generate on a computer sequences of statistically independent random numbers with a normal distribution law. Let γ_1 and γ_2 be two numbers from a sequence distributed to a uniform law in the interval from 0 to 1. Then the sequences

$$\xi = (-2 \ln \gamma_1)^{1/2} \cos 2\pi\gamma_2, \quad \eta = (-2 \ln \gamma_1)^{1/2} \sin 2\pi\gamma_2$$

will be independent sequences of normally distributed random numbers with zero mean and unit variance, and the sequences

$$x = \sigma\xi + \mu$$

$$y = \sigma\eta + \mu$$

will have a normal distribution with mean μ and variance σ^2.

REFERENCES

1. Wiener, N.: *Cybernetics or Control and Communication in the Animal and the Machine.* New York, 1961.
2. Berg, A.I., *et al.* (eds.): *Upravleniye. Informatsiya, Intellekt* (Control. Information. Intelligence). Mysl, Moscow, 1976.
3. Beer, S.: *Cybernetics and Management.* E.C.I., *London.*
4. Ashby, W.R.: *An Introduction to Cybernetics.* Chapman and Hall, London, 1956.
5. Kuzin, L.T.: *Osnovy kibernetiki* (Fundamentals of cybernetics. Vol. 1. Mathematical foundations of cybernetics). Energiya, Moscow, 1973.
6. Kuzin, L.T.: *Osnovy kibernetiki* (Fundamentals of cybernetics. Vol. 2. Cybernetic models). Energiya, Moscow, 1979.
7. Korchounov, Y.: *Foundements Mathématiques de la Cybernétique.* Mir, Moscow, 1975.
8. Good, H.H. and E.R. Machol: *System Engineering. An Introduction to the Design of Large-Scale Systems.* New York, 1957.
9. Stoll, R.R.: *Sets, Logic and Axiomatic Theories.* Freeman, San Francisco, London.
10. Kemeny, J.C., J.L. Snell, and G.L. Thomson: *Introduction to Finite Mathematics.* 2nd ed. Prentice Hall, Englewood Cliffs, N.J., 1966.
11. Berge, Cl.: *Théorie des Graphes et ses Applications.* Bunod, Paris, 1958.
12. Ore, O.: *Theory of Graphs.* American Mathematical Society, New York, 1962.
13. Ford, L.R. and D.R. Fulkerson: *Flows in Networks.* Princeton Univ. Press, Princeton, N.J., 1962.
14. Kolmogorov, A. and S. Fomine: *Elements de la théorie des founctions et de l'analyse founctionnelle.* Mir, Moscow, 1974.
15. Wolfowitz, J.: *Coding Theorems of Information Theory.* Springer Verlag, Berlin, New York, 1964.
16. Pratt W.K.: *Digital Image Processing.* J. Wiley and Sons, New York, 1978.
17. Tow, Ju. and P. Gonzalez: *Pattern Recognition Principles,* 1974.
18. Lanczos, C.: *Applied Analysis.* Prentice Hall, Princeton, N.J., 1956.
19. Bronshtein, I. and K. Semendyaev: *Manual de matemáticas para ingenieros y estudiants.* Mir, Moscou, 1983.
20. Ventsel, E.: *Théory des probabilités.* Mir, Moscou, 1982.
21. Jahnke, E., F. Ende, and F. Lösch. *Tables of Higher Functions.* McGraw Hill, New York, 1960.
22. Wilks, S. St.: *Mathematical Statistics.* Princeton Univ. Press, Princeton, N.J., 1946.
23. Smirnov, V.N. and I.V. Dunin-Barkovsky: *Kurs teorii veroyatnostei i matematicheskoi statistiki* (Probability theory and mathematical statistics). Nauka, Moscow, 1965.

24. Gnedenko, B., Y. Belyaev, and A. Soloviev: *Méthodes mathématiques en théorie de la fiabilite.* Mir, Moscou, 1972.
25. Zehmann, E.Z.: *Testing statistical hypotheses.* Wiley, New York, London.
26. Adler, Yu., E. Markova, and Yu. Granovsky: *The Design Experiments to Find Optimal Conditions.* Mir, Moscow, 1975.
27. Fisher, R.A.: *The Design of Experiments.* Oliver, Edinburgh, London, 1960.
28. Daniel, C.: *Application of Statistics to Industrial Experimentation.* Wiley, New York, 1976.
29. Williams, E.J.: *Regression Analysis.* Wiley, New York, 1959.
30. Tou, J.T.: *Modern Control Theory.* McGraw-Hill, New York, 1964.
31. Wentzel, E.: *Operations Research. A Fundamental Approach.* Mir, Moscow, 1983.
32. Churchman, C.W., R.L. Ackoff, and E.L. Arnoff: *Introduction to Operations Research.* Wiley, New York, 1957.
33. Boltyanski, V.: *Commande optimale des systéms discrets.* Mir, Moscou, 1976.
34. Tsypkin, Ya. Z.: *Adaptatsiya i obuchenie v avtomaticheskikh sistemakh* (Adaptation and teaching in automatic systems). Nauka, Moscow, 1968.
35. Lancaster, K.: *Mathematical Economics.* Macmillan, New York.
36. Moiseyev, N.N., Yu.P. Ivanilov, and E.M. Stolyarova: *Metody optimizatsii* (Optimization techniques). Nauka, Moscow, 1978.
37. Moiseyev, N.N.: *Matematicheskie zadachi sistemnogo analiza* (Mathematical problems of system analysis). Nauka, Moscow, 1981.
38. Carr, Ch. and Ch. W. Howe: *Quantitative Decision Procedures in Management and Economics.*
39. Hadley, G.: *Nonlinear and Dynamic Programming.* Addison Wesley. Reading, Massachusetts, 1964.
40. Wagner, H.W.: *Principles of Operations Research.* Prentice-Hall, Englewood Cliffs, N.J., 1969.
41. Seaty, T.L.: *Optimization in Integers and Related Extremal Problems.* McGraw-Hill, New York, 1970.
42. Danskin, J.M.: *The Theory of Max-Min.* Springer Verlag, Berlin, Heidelberg, New York, 1967.
43. Kuhn, H.W. and A.W. Tucker (eds.): *Linear Inequalities and Related Systems.* Methuen and Co. London.; Wiley, New York, 1958.
44. Golstein, E. and D. Youdin.: *Problems particuliers de la programmation linéaire.* Mir, Moscou, 1973.
45. Gass, S.L.: *Linear Programming.* McGraw-Hill, New York, 1958.
46. Ackoff, R.L.: *A Concept of Corporate Planning.* Wiley, New York.
47. Yudin, D.B.: *Matematicheskie metody upravleniya v usloviyakh nepolnoi informatsii* (Mathematical methods of control under uncertainty). Sovetskoe Radio, Moscow, 1974.
48. Traub, J.T.: *Iterative Methods for Solution of Equations.* Prentice Hall, Englewood Cliffs, N.J., 1969.
49. Pervozvanskii, A.A.: *Poisk* (Search). Nauka, Moscow, 1970.
50. Elsgoltz, L.: *Differential Equations and the Calculus of Variations.* Mir, Moscow, 1980.
51. Pontryagin, L.S., V.G. Boltyanskii, R.V. Gamkrelidze, and R.V. Mishchenko: *The Mathematical Theory of Optimal Processes.* Wiley Interscience. New York, 1962.
52. Boltyanskii, V.G.: *Matematicheskie metody optimalnogo upravleniya* (Mathematical methods of optimal control). Nauka, Moscow, 1969. Boltyanskii, V.G.: "Sufficient Conditions for Optimality and the Justification of the Dynamic Programming Methods", *SIAM J. Control* 1, 326-361 (1966).
53. Rozonoer, L.I.: *Avtomatika Telemekh.* 20, 1320-1334, 1441-1458, 1561-1578 (1959).
54. Bellman, R.: *Dynamic Programming.* Princeton Univ. Press, Princeton, N.J., 1957.

55. Bellman, R.: *Adaptive Control Processes.* Princeton Univ. Press, Princeton. N.J., 1961.
56. Wentsel, E.S.: *Elementy dinamicheskogo programmirovaniya* (Elements of dynamic programming). Nauka, Moscow, 1964.
57. Blackwell, D. and M.A. Girshick: *Theory of Games and Statistical Decisions.* Wiley, New York, Chapman and Hall, London, 1954.
58. McKinsey, J.C.C.: *Introduction to the Theory of Games.* McGraw-Hill, New York, 1954.
59. Von Neumann, J. and O. Morgenstern: *Theory of Games and Economic Behavior.* Princeton Univ. Press, Princeton, N.J., 1953.
60. Chernoff, H. and L.E. Moses: *Elementary Decision Theory.* Wiley, New York.
61. Middleton, D.: *An Introduction to Statistical Communication Theory.* McGraw-Hill, New York, 1960.
62. Wald, A.: *Sequential Analysis.* Wiley, New York, 1947.
63. Bashrinov A.E. and V.S. Fleishman: *Metody statisticheskogo posledovatelnogo analiza i ikh prilozheniya* (Methods of statistical sequential decision analysis and applications). Sov. Radio, Moscow, 1962.
64. Sage, A.P. and J.L. Melse: *Estimation Theory with Application to Communication and Control.* McGraw-Hill, New York, 1972.
65. Speedy, C.B., R.F. Brown, and G.C. Goodwin: *Control Theory.* Oliver and Boyd, Edinburgh, 1970.
66. Meditch, J.S.: *Stochastic Optimal Linear Estimation and Control.* McGraw-Hill, New York, 1969.
67. Korshunov, Yu.M. and A.I. Bobikov: *Tsifrovye sglazhivayushchie i preobrazuyushchie sistemy* (Digital smoothing and converting systems). Energiya, Moscow, 1969.
68. Bryson, A.E. and Y.C.Ho: *Applied Optimal Control.* Blaisdell Publ. Co., Waltham, Massachusetts, 1969.
69. Machol, R.E. (ed.): *System Engineering. A Handbook.* McGraw Hill, New York.
70. Gradshtein, I.S. and I.M. Ryzhik (eds.): *Tables of Integrals, Series and Products.* Academic Press, New York, 1965.
71. Riordan, J.: *Stochastic Service Systems.* Wiley, New York, 1962.
72. Moiseyev, N.N.: *Chislennye metody v teorii optimalnykh sistem.* (Numerical methods in optimal system theory). Nauka, Moscow, 1971.
73. Tanayev, V.S., V.S. Gordon, and Ya. M. Shafranskii: *Teoriya Raspisanii. Odnostadiinye sistemy* (Scheduling theory. One-Stage systems). Nauka, Moscow, 1984.
74. Ackoff, R.L. and M.W. Sasieni: *Fundamentals of Operations Research,* Wiley, New York, 1968.
75. Biegel, Y.E.: *Production Control. A Quantitative Approach.* Prentice-Hall, Englewood Cliffs, N.J., 1971.
76. Riggs, J.L.: *Production Systems,* Wiley, New York, 1970.
77. Hall, M. Jr.: *Combinatorial Theory.* Blaisdell Publ. Co., Waltham, Mass., 1967.
78. Conway, R.W., W.L. Maxwell, and L.W. Miller: *Theory of Scheduling.* Addison-Wesley, Reading, Mass., 1967.
79. Saaty, T.L.: *Elements of Queueing Theory.* McGraw-Hill, New York, 1961.
80. Sobol, I.: *The Monte Carlo Method.* Mir, Moscow, 1984.
81. Hammersley, J.M. and D.C. Handcomb: *Monte-Carlo Methods.* Methuen, London; Wiley, New York, 1964.

SUBJECT INDEX

Alphabet, 15
Alias pairs, 197

Banach space, 93
Bayesian sequential analysis, 339
Bayes theorem, 130
Binary alphabet, 15
Branch and bound algorithm, 381

Capacity constraint, 69
Cartesian product, 38
Central limit theorem, 171
Chapman-Kolmogorov equation, 164
Chromatic number, 62
Circuit, 56
Code, 83, 84, 85
Communication channels, 14
Complement, 33
Confidence interval, 178
Control, 13, 199
  final state, 280
  optimal, 204, 267
  sequential, 223, 325
  system, 16
Controlled system, 17
Correspondence, 25
Correlation, 147
Cut, 70

Data coding, 14
Decision theory, 314, 336
Defining contrast, 197
Discrete time system, 217
Distribution function, 132, 143
Distributions
  binomial, 155
  chi-square, 175
  exponential, 157
  normal, 135, 149
  Poisson, 156
  posterior, 330
  sample mean, 174
  Student *t*, 180
Dynamic optimization problems, 221

Edge, 55
  adjacent, 58
  capacity of, 69
  incident, 57
  pendant, 58
Eigenvalues, 106
Eigenvectors, 106
Equation of motion, 212

Estimation, 172, 344
  confidence interval, 177
  maximum likelihood, 175
Expectation, 136
Experimental design, 191

Failure rate, 158
Filtering, 344
Fractional replication, 195
Full factorial design, 192

Games
  S-game, 299, 318
  solving, 306
  statistical, 314
  strictly determined, 293
  two-person zero-sum, 287
  values of, 293
Game theory, 285
Gaussian Markov process, 165
Graph, 55
  characteristics, 62
  connected acyclic, 59
  isomorphic, 60
  minimum-weighted path length, 67
  nonoriented, 57
  partial, 56
  simple, 56

Hazard rate, 158

Intersection, 31

Kalman filter, 355, 358
Kuhn-Tucker conditions, 255

Law of large numbers, 171
Least-squares method, 346
Limit theorem, 170
Linear dynamic system, 214
Linear equations, 104
Linear space, 89
  dimension of, 92
  normed, 93

Mapping, 42
Markov chain, 159
Mathematical programming, 219
Matrix, 97
  adjacency, 57

algebra, 99
all-vertex incidence, 57
determinant of, 103
diagonalization of, 108, 111
dispersion, 168
inverse, 104
regular, 104
square, 101
symmetric, 110
transpose of, 101
Maximum flow problem, 71
Maximum likelihood principle, 332
Mean value, 140
Mean time to failure, 158
Median, 134
Message space, 83
Minimax theorem, 303, 305
Modal value, 134
Moments, 139

Norm, 93
Notation
binary coded decimal, 16
octal, 16
Nullity, 62

Objective function, 218
Optimization, 227, 259
Ordering, 37

Partition, 34
Path, 56
Pattern recognition, 87
Probability, 123, 124, 125
conditional, 128
density function, 132
distribution, 125
distribution function, 132, 143
inverse, 131
of compound events, 121
posterior, 127, 330
prior, 126
Programming
dynamic, 267
integer, 251
linear, 235
nonlinear, 252
quadratic, 257

Queueing theory, 386

Radix, 15
Random sample, 169
Regression analysis, 146, 185
Relation, 40
composition of, 41
dominance, 53
equivalence, 52
inverse, 41
order, 53
Reliability, 157

Sample mean, 171
Scheduling and sequencing, 361
Self-loop, 56
Sequential control process, 225
Set
algebra, 29
convex, 114, 118
correspondence of, 24
countable, 26
empty, 23
finite, 22
ordered, 37
projection of, 39
Shortest path problem, 63
Simplex computation, 243
Standard deviation, 141
Standardized variable, 135
Statistics, 168
Steepest descent, 262
Strategies, 286, 288, 293, 306

Trail, 56
Transformation, 97
Transition diagram, 160
Transportation problem, 73
Transport network, 69
Traveling salesman problem, 375

Union, 30
Universe of discourse, 33

Variance, 140
Vector algebra, 90
Vertex, 55
adjacent, 57
covering number, 62
degree of, 58
isolated, 57
pendant, 58
sink, 69
strongly connected, 59

Walk, 56
Weierstrass theorem, 229